#31

COPY

Docket No.: PF-0527-1 DIV

I hereby certify that this correspondence is being deposited with the
United States Postal Service as first class mail in an envelope
addressed to: Commissioner for Patents, Box AF, Washington, D.C. 20231
on 5/20/02
By:
Printed: D. Ellis

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of: Lal et al.

Title: PROSTATE GROWTH-ASSOCIATED MEMBRANE PROTEINS

Serial No.: 09/397,558 Filing Date: September 16, 1999

Examiner: Harris, A. Group Art Unit: 1642

Commissioner for Patents
Box AF
Washington, D.C. 20231

DECLARATION OF LARS MICHAEL FURNESS UNDER 37 C.F.R. § 1.132

I, L. MICHAEL FURNESS, a citizen of the United Kingdom, residing at 2 Brookside, Exning, Newmarket, United Kingdom, declare that:

1. I was employed by Incyte Genomics, Inc. (hereinafter "Incyte") as a Director of Pharmacogenomics until December 31, 2001. I am currently under contract to be a Consultant to Incyte.

2. In 1984, I received a B.Sc.(Hons) in Biomolecular Science (Biophysics and Biochemistry)

from Portsmouth Polytechnic.

From 1985-1987 I was at the School of Pharmacy in London, United Kingdom. during which time I analyzed lipid methyltransferase enzymes using a variety of protein analysis methods, including one-dimensional (1D) and two-dimensional (2D) gel electrophoresis, HPLC, and a variety of enzymatic assay systems.

I then worked in the Protein Structure group at the National Institute for Medical Research until 1989, setting up core facilities for nucleic acid synthesis and sequencing, as well as assisting in programs on protein kinase C inhibitors.

After a year at Perkin Elmer-Applied Biosystems as a technical specialist, I worked at the Imperial Cancer Research Fund between 1990-1992, on a Eureka-funded program collaborating with Amersham Pharmacia in the United Kingdom and CEPH (Centre d'Etude du Polymorphisme Humaine) in Paris, France, to develop novel nucleic acid purification and characterization methods.

In 1992, I moved to Pfizer Central Research in the United Kingdom, where I stayed until 1998, initially setting up core DNA sequencing and then a DNA arraying facility for gene expression analysis in 1993. My work also included bioinformatics and I was responsible for the support of all Pfizer neuroscience programs in the United Kingdom. This then led me into carrying out detailed bioinformatics and wet lab work on the sodium channels, including antibody generation, Western and Northern analyses, PCR, tissue distribution studies, and sequence analyses on novel sequences identified.

In 1998, I moved to Incyte to work in the Pharmacogenomics group, looking at the application of genomics and proteomics to the pharmaceutical industry. In 1999, I was appointed Director of the LifeExpress Lead Program which used microarray and protein expression data to identify pharmacologically and toxicologically relevant mechanisms to assist in improved drug design and development.

On December 12, 2001, I founded Nuomics Consulting, Ltd., in Exning, UK, where I am currently employed as Managing Director. Nuomics Consulting, Ltd. provides expert technical knowledge and advice to businesses in the areas of genomics, proteomics, pharmacogenomics, toxicogenomics, and chemogenomics.

3. I have reviewed the specification of a United States patent application that I understand was filed on September 16, 1999 in the names of Preeti Lal et al. and was assigned Serial No. 09/397,558 (hereinafter "the Lal '558 application"). Furthermore, I understand that this United States patent application was a divisional application of, and claimed priority to, United States patent application Serial No. 09/083,521, filed on May 22, 1998 (hereinafter "the Lal '521 application"), having the identical specification. My remarks herein will therefore be directed to the Lal '521 patent application, and May 22, 1998, as the relevant date of filing. In broad overview, the Lal '521 specification pertains to certain nucleotide and amino acid sequences and their use in a number of applications, including gene and protein expression monitoring applications that are useful in connection with (a) developing drugs (e.g., for the treatment of cancer), and (b) monitoring the activity of drugs for purposes relating to evaluating their efficacy and toxicity.

4. I understand that (a) the Lal '558 application contains claims that are directed to isolated polypeptides having either of the sequences shown as SEQ ID NO:1 and SEQ ID NO:2 (hereinafter "the SEQ ID NO:1 and SEQ ID NO:2 polypeptides"), and (b) the Patent Examiner has rejected those claims on the grounds that the specification of the Lal '558 application does not disclose a specific and substantial asserted utility or a well established utility for the claimed SEQ ID NO:1 and SEQ ID NO:2 polypeptides. I further understand that whether or not a patent specification discloses a specific and substantial asserted utility or a well established utility for its claimed subject matter is properly determined from the perspective of a person skilled in the art to which the specification pertains at the time the patent application was filed. In addition, I understand that a specific and substantial asserted utility or a well established utility under the patent laws must be a "real-world" utility.

5. I have been asked (a) to consider with a view to reaching a conclusion (or conclusions) as to whether or not I agree with the Patent Examiner's position that the Lal '558 application and its parent, the Lal '521 application, do not disclose a specific and substantial "real-world" utility for the claimed SEQ ID NO:1 and SEQ ID NO:2 polypeptides, and (b) to state and explain the bases for any conclusions I reach. I have been informed that, in connection with my considerations, I should determine whether or not a person skilled in the art to which the Lal '521 application pertains on May

22, 1998, would have concluded that the Lal '521 application disclosed, for the benefit of the public, a specific beneficial use of the SEQ ID NO:1 and SEQ ID NO:2 polypeptides in their then available and disclosed forms. I have also been informed that, with respect to the "real-world" utility requirement, the Patent and Trademark Office instructs its Patent Examiners in Section 2107 of the Manual of Patent Examining Procedure, under the heading "I. 'Real-World Value' Requirement":

> "Many research tools such as gas chromatographs, screening assays, and nucleotide sequencing techniques have a clear, specific and unquestionable utility (e.g., they are useful in analyzing compounds). An assessment that focuses on whether an invention is useful only in a research setting thus does not address whether the specific invention is in fact 'useful' in a patent sense. Instead, Office personnel must distinguish between inventions that have a specifically identified substantial utility and inventions whose asserted utility requires further research to identify or reasonably confirm."

6. I have considered the matters set forth in paragraph 5 of this Declaration and have concluded that, contrary to the position I understand the Patent Examiner has taken, the specification of the Lal '521 patent application disclosed to a person skilled in the art at the time of its filing a number of specific and substantial real-world utilities for the claimed SEQ ID NO:1 and SEQ ID NO:2 polypeptides. More specifically, persons skilled in the art on May 22, 1998, would have understood the Lal '521 application to disclose the use of the SEQ ID NO:1 and SEQ ID NO:2 polypeptides as research tools in a number of gene and protein expression monitoring applications that were well-known at that time to be useful in connection with the development of drugs and the monitoring of the activity of such drugs. I explain the bases for reaching my conclusion in this regard in paragraphs 7-13 below.

7. In reaching the conclusion stated in paragraph 6 of this Declaration, I considered (a) the specification of the Lal '521 application, and (b) a number of published articles and patent documents that evidence gene and protein expression monitoring techniques that were well-known before the

May 22, 1998 filing date of the Lal '521 application. The published articles and patent documents I considered are:

(a) Anderson, N.L., Esquer-Blasco, R., Hofmann, J.-P., Anderson, N.G., A Two-Dimensional Gel Database of Rat Liver Proteins Useful in Gene Regulation and Drug Effects Studies, Electrophoresis, 12, 907-930 (1991) (hereinafter "the Anderson 1991 article") (copy annexed at Tab A);

(b) Anderson, N.L., Esquer-Blasco, R., Hofmann, J.-P., Mehues, L., Raymackers, J., Steiner, S., Witzmann, F., Anderson, N.G., An Updated Two-Dimensional Gel Database of Rat Liver Proteins Useful in Gene Regulation and Drug Effect Studies, Electrophoresis, 16, 1977-1991 (1995) (hereinafter "the Anderson 1995 article") (copy annexed at Tab B);

(c) Wilkins, M.R., Sanchez, J.-C., Gooley, A.A., Appel, R.D., Humphrey-Smith, I., Hochstrasser, D.F., Williams, K.L., Progress with Proteome Projects: Why all Proteins Expressed by a Genome Should be Identified and How To Do It, Biotechnology and Genetic Engineering Reviews, 13, 19-50 (1995) (hereinafter "the Wilkins article") (copy annexed at Tab C);

(d) Celis, J.E., Rasmussen, H.H., Leffers, H., Madsen, P., Honore, B., Gesser, B., Dejgaard, K., Vandekerckhove, J., Human Cellular Protein Patterns and their Link to Genome DNA Sequence Data: Usefulness of Two-Dimentional Gel Electrophoresis and Microsequencing, FASEB Journal, 5, 2200-2208 (1991) (hereinafter "the Celis article") (copy annexed at Tab D);

(e) Franzen, B., Linder, S., Okuzawa, K., Kato, H., Auer, G., Nonenzymatic Extraction of Cells from Clinical Tumor Material for Analysis of Gene Expression by Two-Dimensional Polyacrylamide Gel Electrophoresis, Electrophoresis, 14, 1045-1053 (1993) (hereinafter "the Franzen article") (copy annexed at Tab E);

(f) Bjellqvist, B., Basse, B., Olsen, E., Celis, J.E., Reference Points for Comparisons of Two-Dimensional Maps of Proteins from Different Human Cell Types Defined in a pH Scale Where Isoelectric Points Correlate with Polypeptide Compositions, Electrophoresis, 15, 529-539 (1994) (hereinafter "the Bjellqvist article") (copy annexed at Tab F); and

(g) Large Scale Biology Company Info; LSB and LSP Information; from http://www.lsbc.com (2001) (copy annexed at Tab G).

8. Many of the published articles I considered (i.e., at least items (a)-(f) identified in paragraph 7) relate to the development of protein two-dimensional gel electrophoretic techniques for use in gene and protein expression monitoring applications in drug development and toxicology. As I will discuss below, a person skilled in the art who read the Lal '521 application on May 22, 1998 would have understood that application to disclose the SEQ ID NO:1 and SEQ ID NO:2 polypeptides to be useful for a number of gene and protein expression monitoring applications, e.g., in the use of two-dimensional polyacrylamide gel electrophoresis and western blot analysis of tissue samples in drug development and in toxicity testing.

9. Turning more specifically to the Lal '521 specification, the SEQ ID NO:1 and SEQ ID NO:2 polypeptides are shown at pages 51-53 as two of seven sequences under the heading "Sequence Listing." The Lal '521 specification specifically teaches that the "invention features substantially purified polypeptides, prostate growth-associated membrane proteins, referred to collectively as 'PGAMP' and individually as 'PGAMP-1' and 'PGAMP-2.' In one aspect, the invention provides a substantially purified polypeptide comprising an amino acid sequence selected from the group consisting of SEQ ID NO:1, SEQ ID NO:2, a fragment of SEQ ID NO:1, and a fragment of SEQ ID NO:2." (Lal '521 application at page 3, lines 5-9, as amended). With respect to SEQ ID NO:1, the Lal '521 specification teaches that (a) the identity of the SEQ ID NO:1 polypeptide was determined from a "prostate cDNA library", (b) the SEQ ID NO:1 polypeptide is the human prostate growth-associated membrane protein referred to as "PGAMP-1" and is encoded by SEQ ID NO:3, and (c) northern analysis shows that PGAMP-1 is expressed "in various libraries, at least 72% of which are immortalized or cancerous and at least 18% of which invlove immune response. Of particular note is the expression of PGAMP-1 in cancerous or hyperplastic prostate (48%) and breast (7%)" tissues and therefore PGAMP-1 "appears to play a role in neoplastic and reproductive disorders" (Lal '521 application at page 13, lines 27-32; page 14, lines 10-13; and page 25, lines 15-17). With respect to SEQ ID NO:2, the Lal '521 specification teaches that (a) the identity of the SEQ ID NO:2 polypeptide was determined from a "breast cDNA library", (b) the SEQ ID NO:2 polypeptide is the human prostate growth-associated membrane protein referred to as "PGAMP-2" and is encoded by

SEQ ID NO:4, and (c) northern analysis shows that PGAMP-2 is expressed "in various libraries, at least 76% of which are immortalized or cancerous and at least 18% of which invlove immune response. Of particular note is the expression of PGAMP-2 in cancerous or hyperplastic prostate (28%) and breast (10%)" tissues and therefore PGAMP-2 "appears to play a role in neoplastic and reproductive disorders" (Lal '521 application at page 14, lines 14-19; page 15, lines 4-8; and page 25, lines 20-22).

The Lal '521 application discusses a number of uses of the SEQ ID NO:1 and SEQ ID NO:2 polypeptides in addition to their use in gene and protein expression monitoring applications. I have not fully evaluated these additional uses in connection with the preparation of this Declaration and do not express any views in this Declaration regarding whether or not the Lal '521 specification discloses these additional uses to be substantial, specific and credible real-world utilities of the SEQ ID NO:1 and SEQ ID NO:2 polypeptides. Consequently, my discussion in this Declaration concerning the Lal '521 application focuses on the portions of the application that relate to the use of the SEQ ID NO:1 and SEQ ID NO:2 polypeptides in gene and protein expression monitoring applications.

10. The Lal '521 application discloses that the polynucleotide sequences disclosed therein, including the polynucleotides encoding the SEQ ID NO:1 and SEQ ID NO:2 polypeptides, are useful as probes in chip based technologies. It further teaches that the chip based technologies can be used "for the detection and/or quantification of nucleic acid or protein" (Lal '521 application at page 23, lines 5-8).

The Lal '521 application also discloses that the SEQ ID NO:1 and SEQ ID NO:2 polypeptides are useful in other protein expression detection technologies. The Lal '521 application states that "[I]mmunological methods for detecting and measuring the expression of PGAMP using either specific polyclonal or monoclonal antibodies are known in the art. Examples of such techniques include enzyme-linked immunosorbent assays (ELISAs), radioimmunoassays (RIAs), and fluorescence activated cell sorting (FACS)" (Lal '521 application at page 23, lines 9-12). Furthermore, the Lal '521 application discloses that "[a] variety of protocols for measuring PGAMP, including ELISAs, RIAs, and FACS, are known in the art and provide a basis for diagnosing altered or abnormal levels of PGAMP expression. Normal or standard values for PGAMP expression are established by combining body fluids or cell extracts taken from normal mammalian subjects, preferably human, with antibody to PGAMP under conditions suitable for complex formation" (Lal '521 application at page 34, lines 2-6).

In addition, at the time of filing the Lal '521 application, it was well known in the art that "gene" and protein expression analyses also included two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) technologies, which were developed during the 1980s, as exemplified by the Anderson 1991 and 1995 articles (Tab A and Tab B). The Anderson 1991 article teaches that a 2-D PAGE map has been used to connect and compare hundreds of 2-D gels of rat liver samples from a variety of studies including regulation of protein expression by various drugs and toxic agents (Tab A at p. 907). The Anderson 1991 article teaches an empirically-determined standard curve fitted to a series of identified proteins based upon amino acid chain length, and how that standard curve can be used in protein expression analysis (Tab A at p. 911). The Anderson 1991 article teaches that "there is a long-term need for a comprehensive database of liver proteins" (Tab A at p. 912).

The Wilkins article is one of a number of documents that were published prior to the May 22, 1998 filing date of the Lal '521 application that describes the use of the 2-D PAGE technology in a wide range of gene and protein expression monitoring applications, including monitoring and analyzing protein expression patterns in human cancer, human serum plasma proteins, and in rodent liver following exposure to toxins. In view of the Lal '521 application, the Wilkins article, and other related pre-May 1998 publications, persons skilled in the art on May 22, 1998 clearly would have understood the Lal '521 application to disclose the SEQ ID NO:1 and SEQ ID NO:2 polypeptides to be useful in 2-D PAGE analyses for the development of new drugs and for monitoring the activities of drugs for such purposes as evaluating their efficacy and toxicity, as explained more fully in paragraph 12 below.

With specific reference to toxicity evaluations, those of skill in the art who were working on drug development in May 1998 (and for many years prior to May 1998) without any doubt appreciated that the toxicity (or lack of toxicity) of any proposed drug they were working on was one of the most important criteria to be considered and evaluated in connection with the development of the drug. They would have understood at that time that good drugs are not only potent, they are specific. This means that they have strong effects on a specific biological target and minimal effects on all other biological targets. Ascertaining that a candidate drug affects its intended target, and identifying undesirable secondary effects (i.e., toxic side effects), had been for many years among the main challenges in developing new drugs. The ability to determine which genes are positively affected by a given drug, coupled with the ability to quickly and at the earliest time possible in the drug development process identify drugs that are likely to be toxic because of their undesirable secondary effects, have enormous value in improving the efficiency of the drug discovery process, and are an important and essential part of the development of any new drug. In fact, the desire to identify and understand toxicological effects using the experimental assays described above led Dr Leigh Anderson to found the Large Scale Biology Corporation in 1987, in order to pursue commercial development of the 2-D electrophoretic protein mapping technology he had developed. In addition, the company focused on toxicological effects on the proteome as clearly demonstrated by its goals and by its senior management credentials described in company documents (see Tab G at pp. 1, 3, and 5).

Accordingly, the teachings in the Lal '521 application, in particular regarding use of the SEQ ID NO:1 and SEQ ID NO:2 polypeptides in differential gene and protein expression analysis (2-D PAGE maps) and in the development and the monitoring of the activities of drugs, clearly includes toxicity studies, and persons skilled in the art who read the Lal '521 application on May 22, 1998 would have understood that to be so.

11. As previously discussed (*supra*, paragraphs 7 and 8), in the mid-1980s the several publications annexed to this Declaration at Tabs A through F evidence information that was available to the public regarding two-dimensional polyacrylamide gel electrophoresis technology and its uses in drug discovery and toxicology testing before the May 22, 1998 filing date of the Lal '521 application. In particular the Celis article stated that "protein databases are expected to foster a variety of biological information... -- among others, ... drug development and testing" (See Tab D, p. 2200, second column). The Franzen article shows that 2-D PAGE maps were used to identify proteins in clinical tumor material (See Tab E). The Lal '521 application clearly discloses that expression of PGAMP-1 and/or PGAMP-2 is associated with immortalized cell lines, cancerous and

hyperplastic prostate and breast tissue, and with the immune response (Lal '521 application at page 14, lines 10-13; page 15, lines 4-8; and page 25, lines 15-16 and 20-21). The Bjellqvist article showed that a protein may be identified accurately by its positional coordinates, namely molecular mass and isoelectric point (See Tab F). The Lal '521 application clearly disclosed SEQ ID NO:1 and SEQ ID NO:2 from which it would have been routine for one of skill in the art to predict both the molecular mass and the isoelectric point using algorithms well known in the art at the time of filing.

12. A person skilled in the art on May 22, 1998 who read the Lal '521 application, would understand that application to disclose the SEQ ID NO:1 and SEQ ID NO:2 polypeptides to be highly useful in analysis of differential expression of proteins. For example, the specification of the Lal '521 application would have led a person skilled in the art in May 1998, who was using protein expression monitoring in connection with developing new drugs for the treatment of a neoplastic or reproductive disorder to conclude that a 2-D PAGE map that used the substantially purified SEQ ID NO:1 and SEQ ID NO:2 polypeptides would be a highly useful tool and to request specifically that any 2-D PAGE map that was being used for such purposes utilize the SEQ ID NO:1 and/or SEQ ID NO:2 polypeptides. Expressed proteins are useful for 2-D PAGE analysis in toxicology expression studies for a variety of reasons, particularly for purposes relating to providing controls for the 2-D PAGE analysis, and for identifying sequence or post-translational variants of the expressed sequences in response to exogenous compounds. Persons skilled in the art would appreciate that a 2-D PAGE map that utilized the SEQ ID NO:1 and SEQ ID NO:2 polypeptide sequences would be a more useful tool than a 2-D PAGE map that did not utilize these protein sequences in connection with conducting protein expression monitoring studies on proposed (or actual) drugs for treating neoplastic and reproductive disorders for such purposes as evaluating their efficacy and toxicity.

I discuss in more detail in items (a)-(b) below a number of reasons why a person skilled in the art, who read the Lal '521 specification in May 1998, would have concluded based on that specification and the state of the art at that time, that the SEQ ID NO:1 and SEQ ID NO:2 polypeptides would be highly useful tools for analysis of a 2-D PAGE map for evaluating the efficacy and toxicity of proposed drugs for neoplastic and reproductive disorders by means of 2-D PAGE maps, as well as for other evaluations.

(a) The Lal '521 specification contains a number of teachings that would lead persons skilled in the art on May 22, 1998 to conclude that a 2-D PAGE map that utilized the substantially purified SEQ ID NO:1 and/or SEQ ID NO:2 polypeptides would be a more useful tool for gene and protein expression monitoring applications relating to drugs for treating neoplastic and reproductive disorders than a 2-D PAGE map that did not use the SEQ ID NO:1 and/or SEQ ID NO:2 polypeptides. Among other things, the Lal '521 specification teaches that (i) the identity of the SEQ ID NO:1 polypeptide was determined from a prostate cDNA library, (ii) the SEQ ID NO:1 polypeptide is the prostate growth-associated membrane protein referred to as PGAMP-1, and (iii) PGAMP-1 is expressed in various libraries derived from immortalized and cancerous tissues, cancerous or hyperplastic prostate and breast tissues, and tissues involved in the immune response, and, therefore, PGAMP-1 expression is "associated with neoplastic and reproductive disorders" (Lal '521 application at page 13, lines 27-32; page 14, lines 10-13; and page 25, lines 15-17; see paragraph 9, *supra*). Furthermore, the Lal '521 specification teaches that (i) the identity of the SEQ ID NO:2 polypeptide was determined from a breast cDNA library, (ii) the SEQ ID NO:2 polypeptide is the prostate growth-associated membrane protein referred to as PGAMP-2, and (iii) PGAMP-2 is expressed in various libraries derived from immortalized and cancerous tissues, cancerous or hyperplastic prostate and breast tissues, and tissues involved in the immune response, and, therefore, PGAMP-2 expression is "associated with neoplastic and reproductive disorders" (Lal '521 application at page 14, lines 14-19; page 15, lines 4-8; and page 25, lines 20-22; see paragraph 9, *supra*). The substantially purified SEQ ID NO:1 and SEQ ID NO:2 polypeptides could, therefore, be used as controls to more accurately gauge the expression of PGAMP in a sample, and consequently more accurately gauge the effect of a toxicant on expression of the gene.

Moreover, the Lal '521 specification teaches that SEQ ID NO:1 and SEQ ID NO:2 share chemical and structural homology with known tumor-associated antigens. PGAMP-1 shares chemical and structural homology with rat heat-stable antigen CD4. These polypeptides share 21% identity and two potential transmembrane domains (Lal '521 application at page 14, lines 6-8; and Figure 1). In addition, PGAMP-1 has chemical similarity with CD44 antigen precursor (Lal '521 application at page 14, lines 1-5). PGAMP-2 shares chemical and structural homology with human prostate-specific antigen and a fragment of the mouse apoptosis-associated tyrosine kinase, sharing

18% and 17% identity, respectively (Lal '521 application at page 14, lines 29-33; and Figures 2A. 2B, and 2C). In addition, all three of these proteins share six potential transmembrane regions and a potential signal peptide, and PGAMP-2 and human prostate-specific antigen have similar isoelectric points (Lal '521 application at page 15, lines 1-2).

(b) Persons skilled in the art on May 22, 1998 would have appreciated (i) that the protein expression monitoring results obtained using a 2-D PAGE map that utilized the SEQ ID NO:1 and/or SEQ ID NO:2 polypeptides would vary, depending on the particular drug being evaluated, and (ii) that such varying results would occur both with respect to the results obtained from the SEQ ID NO:1 and/or SEQ ID NO:2 polypeptides and from the 2-D PAGE map as a whole (including all its other individual proteins). These kinds of varying results, depending on the identity of the drug being tested, in no way detract from my conclusion that persons skilled in the art on May 22, 1998, having read the Lal '521 specification, would specifically request that any 2-D PAGE map that was being used for conducting protein expression monitoring studies on drugs for treating neoplastic and reproductive disorders (*e.g.*, a toxicology study or any efficacy study of the type that typically takes place in connection with the development of a drug) utilize the SEQ ID NO:1 and/or SEQ ID NO:2 polypeptides. Persons skilled in the art on May 22, 1998 would have wanted their 2-D PAGE map to utilize the SEQ ID NO:1 and/or SEQ ID NO:2 polypeptides because a 2-D PAGE map that utilized these polypeptides (as compared to one that did not) would provide more useful results in the kind of gene and protein expression monitoring studies using 2-D PAGE maps that persons skilled in the art have been doing since well prior to May 22, 1998.

The foregoing is not intended to be an all-inclusive explanation of all my reasons for reaching the conclusions stated in this paragraph 12, and in paragraph 6, *supra*. In my view, however, it provides more than sufficient reasons to justify my conclusions stated in paragraph 6 of this Declaration regarding the Lal '521 application disclosing to persons skilled in the art at the time of its filing substantial, specific and credible real-world utilities for the SEQ ID NO:1 and SEQ ID NO:2 polypeptides.

13. Also pertinent to my considerations underlying this Declaration is the fact that the Lal '521 disclosure regarding the uses of the SEQ ID NO:1 and SEQ ID NO:2 polypeptides for protein

expression monitoring applications is not limited to the use of these proteins in 2-D PAGE maps. For one thing, the Lal '521 disclosure regarding the technique used in gene and protein expression monitoring applications is broad (Lal '521 application at, e.g., page 23, lines 3 to 31; and page 34, lines 2-10).

In addition, the Lal '521 specification repeatedly teaches that the proteins described therein (including the SEQ ID NO:1 and SEQ ID NO:2 polypeptides) may desirably be used in any of a number of long established "standard" techniques, such as ELISA or western blot analysis, for conducting protein expression monitoring studies. See, e.g.:

(a) Lal '521 application at p. 23, lines 9-12 ("Immunological methods for detecting and measuring the expression of PGAMP using either specific polyclonal or monoclonal antibodies are known in the art. Examples of such techniques include enzyme-linked immunosorbent assays (ELISAs), radioimmunoassays (RIAs), and fluorescence activated cell sorting (FACS)"); and

(b) Lal '521 application at p. 34, lines 2-10 ("A variety of protocols for measuring PGAMP, including ELISAs, RIAs, and FACS, are known in the art and provide a basis for diagnosing altered or abnormal levels of PGAMP expression. Normal or standard values for PGAMP expression are established by combining body fluids or cell extracts taken from normal mammalian subjects, preferably human, with antibody to PGAMP under conditions suitable for complex formation[.] The amount of standard complex formation may be quantified by various methods, preferably by photometric means. Quantities of PGAMP expressed in subject, control, and disease samples from biopsied tissues are compared with the standard values. Deviation between standard and subject values establishes the parameters for diagnosing disease").

Thus, a person skilled in the art on May 22, 1998, who read the Lal '521 specification, would have routinely and readily appreciated that the SEQ ID NO:1 and SEQ ID NO:2 polypeptides, disclosed therein, would be useful to conduct gene and protein expression monitoring analyses using 2-D PAGE mapping or western blot analysis or any of the other traditional membrane-based protein expression monitoring techniques that were known and in common use many years prior to the filing of the Lal '521 application. For example, a person skilled in the art in May 1998 would have routinely and readily appreciated that the SEQ ID NO:1 and SEQ ID NO:2 polypeptides would be useful tools in conducting protein expression analyses, using the 2-D PAGE mapping or western analysis techniques, in furtherance of (a) the development of drugs for the treatment of neoplastic and reproductive disorders, and (b) analyses of the efficacy and toxicity of such drugs.

14. I declare further that all statements made herein of my own knowledge are true and that all statements made herein on information and belief are believed to be true; and further, that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, and that willful false statements may jeopardize the validity of this application and any patent issuing thereon.

L. Michael Furness, B.Sc.

Signed at Exning, United Kingdom
this 8th day of February, 2002.

A

Leigh Anderson[1]
Jeff Seilhamer[2]

[1]Large Scale Biology Corporation, Rockville, MD, USA
[2]Incyte Pharmaceuticals, Palo Alto, CA, USA

A comparison of selected mRNA and protein abundances in human liver

In order to obtain an estimate of the overall level of correlation between mRNA and protein abundances for a well-characterized pharmaceutically relevant biological system, we have analyzed human liver by quantitative two-dimensional electrophoresis (for protein abundances) and by Transcript Image methodology (for mRNA abundances). Incyte's LifeSeq® database was searched for expressed sequence tag (EST) sequences corresponding to a series of 23 proteins identified on 2-D maps in the Large Scale Biology (LSB) Molecular Anatomy™ database, resulting in estimated abundances for 19 messages (4 were undetected) among 7926 liver clones sequenced. A correlation coefficient of 0.48 was obtained between the mRNA and protein abundances determined by the two approaches, suggesting that post-transcriptional regulation of gene expression is a frequent phenomenon in higher organisms. A comparison with published data (Kawamoto, S., *et al.*, *Gene* 1996, *174*, 151–158) on the abundances of liver mRNAs for plasma proteins (secreted by the liver) suggests that higher abundance messages are strongly enriched in secreted sequences. Our data confirms this: of the 50 most abundant liver mRNAs, 29 coded for secreted proteins, while none of the 50 most abundant proteins appeared to be secreted products (although four plasma and red blood cell proteins were present in this group as contaminants from tissue blood).

1 Introduction

The control of gene expression is achieved by a series of complex mechanisms which can be divided into two basic phases. The first phase, which involves the processing of sequence information from DNA, through transcription, RNA splicing, and transport through the nuclear membrane to yield a mature mRNA, has been relatively well characterized for many genes through nucleic acid sequencing approaches. The second phase, involving translation into protein (dependent on mRNA translatability), folding, assembly into multimers, transport to an appropriate subcellular location, post-translational modifications, and final destruction, has been less comprehensively characterized. Both phases are likely to contain important control points associated with gene regulation underlying differentiation, disease processes and drug effects. For a variety of reasons, it would be useful to know the extent to which mRNA abundances are predictive of corresponding protein abundances. A series of powerful methodologies, including Transcript Imaging [1], SAGE [2], differential display [3] and array hybridization [4–6], have been developed to detect and in some cases quantitate differences in mRNA composition between different samples. In parallel, high resolution protein mapping systems, based on two-dimensional (2-D) electrophoresis [7], have been employed to build quantitative databases describing gene expression at the protein level [8–11]. By combining these approaches, it is possible for the first time to examine both levels at which gene expression is controlled, and thereby to develop a global understanding of gene expression control.

To date, we are aware of surprisingly little published work on the overall relationship of message and protein abundance, with the exception of a recent study by Kawamoto *et al.* [12], comparing mRNA levels obtained for plasma protein genes by transcript image methodology with the abundances of the corresponding plasma proteins in circulation. This report appeared to show a strong correlation between mRNA and protein abundance, based on data for nine human gene products. It seemed likely, however, that such secreted proteins constitute a special case, since they are rapidly delivered from the cell of synthesis to the plasma compartment, where many of the mechanisms that regulate cellular protein abundance are presumably absent. We therefore decided to compare mRNA and protein levels for a larger series of cellular molecules in order to see whether a simple relationship exists between mRNA and protein abundance for this class, and to see whether mRNAs for major cellular proteins are generally more or less abundant than those for major secreted products.

2 Materials and methods

Samples for 2-D electrophoresis were prepared by rapidly mixing a frozen powder of human liver (prepared and stored at liquid nitrogen temperature in the National Biomonitoring Specimen Bank at the US National Institute of Standards and Technology) with an 8-fold excess of 9 M urea, 2% NP-40, 1% mercaptoethanol and 2% carrier ampholytes (LKB 9–11). Ten µL of the resulting sample was analyzed using the Iso-DALT 2-D electrophoresis system, and the gels stained with colloidal Coomassie Brilliant Blue (CBB) G-250 as previously described [13–16]. Each stained slab gel was digitized in red light at 134 µm resolution using an Eikonix 1412 scanner and the digitized gel images pro-

Correspondence: Dr. Leigh Anderson, Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338 USA (Tel: +301-424-5989; Fax: +301-762-4892; email: leigh@lsbc.com)

Nonstandard abbreviation: CBB, Coomassie Brilliant Blue

Keywords: Messenger RNA / Two-dimensional polyacrylamide gel electrophoresis / Transcript image / Liver / Regulation

0173-0835/97/0304-0533 $17.50+.50/0

Table 1. Protein and mRNA abundances in human liver reported for 23 selected molecules

Protein name	Protein	Average protein abundance	Protein standard deviation	Average protein abundance (%)	Number of clones (BLAST)	Average message abundance
Carbamyl phosphate synthase	CPS	101475	12379	2.83	11	0.139
Actin beta	ACTB	50345	17793	1.41	15	0.189%
Heat shock protein 60	HSP60	37656	1939	1.05	3	0.038%
Protein disulfide isomerase	PDI	31260	1942	0.87	2	0.025%
78 KD glucose regulated protein / BIP	BIP	31050	1993	0.87	1	0.013%
Calreticulin	CRTC	30491	2076	0.85	3	0.038%
F1 ATPase beta	F1ATPB	29529	1275	0.82	3	0.038%
Actin gamma	ACTG	23316	9012	0.65	17	0.215%
Heat shock cognate 70	HSC70	21647	908	0.60	1	0.013%
Cytochrome B5	CYB5	18776	1656	0.52	7	0.088
Endoplasmin	ENPL	17817	5829	0.50	5	0.063%
75 KD glucose regulated protein	GR75	16380	1821	0.46	1	0.013%
Pyruvate carboxylase	PYVC	14655	1930	0.41	0	Not detected
Heat shock protein 70	HSP70	8629	1565	0.24	1	0.013%
Tubulin beta 1	TBB1	7125	1472	0.20	3	0.038%
Vimentin	VIME	6269	952	0.18	0	not detected
Tropomyosin	TPM	4090	600	0.11	1	0.013%
NADPH cytochrome P-450 reductase	NP450R	3303	1319	0.09	0	Not detected
Tubulin alpha 1	TBA1	3097	1409	0.09	5	0.063%
Heat shock protein 90	HSP90	2740	597	0.08	2	0.025%
Cytochrome oxidase II (mit encoded)	COX-II	2384	651	0.07	0	Not measured
Laminin receptor	LAMR	1531	602	0.04	4	0.050%
Lamin B	LAMB	1454	371	0.04	2	0.025%

a) Protein abundance is given in pixel-gray levels (the integrated CBB optical density of the appropriate spot or spots on a 2-D gel), where multiple spots comprising a single gene product have been summed. Messenger RNA measurements are given as a percentage of the total number of clones sequenced in the relevant transcript images.

cessed using the Kepler* software system (Large Scale Biology) to give protein abundances in terms of pixel × gray-level values, as well as group average abundances and standard deviations over a set of seven male human livers. Relative abundances were computed by dividing individual average abundances by the average total abundance of the proteins resolved on the gels. A series of proteins was identified on these gels based on close homology with identified rodent liver spots and on identifications published by Hughes *et al.* [17]. Total cellular RNA was extracted from samples of human liver tissue by the method of Chirgwin *et al.* [18], and poly-A+ RNA was prepared by hybridization to oligo-dT cellulose. Five µg of poly-A+ RNA was used to construct a cDNA library using the Gubler and Hoffman method [19] in bacteriophage-lambda UNIZap™ (Stratagene Inc., La Jolla, CA). The library was converted to plasmid DNA by bulk excision, and individual colonies were selected for DNA template preps. The templates were sequenced enzymatically (Sanger *et al.* [20]) on an ABI 373 automated DNA sequencer. Templates considered sequenced sucessfully contained > 230 bases of cDNA insert sequence after removal of repetitive and low information sequences, > 90% base call accuracy, and were not of mitochondrial, vector or host origin. Resulting DNA sequences were analyzed using the BLAST program for similarity with other known primate, mammalian, and subsequently all divisions of GenBank. Similarity data was stored and tabulated in the LifeSeq™ software (Incyte, Palo Alto, CA), from which relative fractions of specific gene products present within the starting RNA prep were calculated as follows: % abundance = # clones representing each gene / total # of genes sampled *100. A total of 7925 clones were sequenced from liver obtained from two individuals: one male (5054 clones) and one female (2871 clones). Data from Table 1 of Kawamoto *et al.* [12], was replotted using protein abundances for human plasma proteins taken as mean values of the range presented in reference [21]. An error in the abundance of the haptoglobin α1s polypeptide (which was assumed in [12] to account for the entire abundance of the haptoglobin $\alpha_2\beta_2$ tetramer) was corrected.

3 Results

Protein and mRNA abundance data were collected for a set of gene products identified on 2-D gels (Table 1). Standard deviations of the protein measurements across six individual livers were relatively low, averaging 19% of the mean abundance. Of the 23 selected proteins, mRNAs for 19 were detected in human liver transcript images. Of these 19, five were represented by 1 clone, three by 2 clones, four by 3 clones, and the rest by between 4 and 17 clones. Of the four gene products undetected at the mRNA level, one (cytochrome oxidase subunit II: COX-II) was deleted from the Transript Image dataset during standard initial sequence data workup, which removes all mitochondrial sequences. A plot of protein abundance (expressed as integrated Coomassie Blue absorbance averaged over seven individual livers) *versus* mRNA abundance (expressed as per-

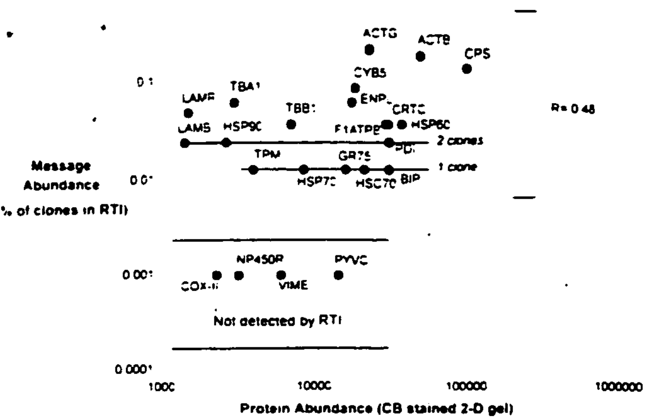


Figure 1. A log-log plot of the abundances of each of 23 gene products at the protein level (*X*-axis) and mRNA level (*Y*-axis). Four proteins for which mRNA measurements were not available – three for which no clones were detected, and one intenionally deleted from the RTI dataset (COX-II) – are shown boxed at the lower left, with correct relative protein abundances. The Pearson product moment correlation coefficient between the two sets of 19 valid measurements is 0.48. Each measurement is labeled with a code whose identity is shown in Table 1.

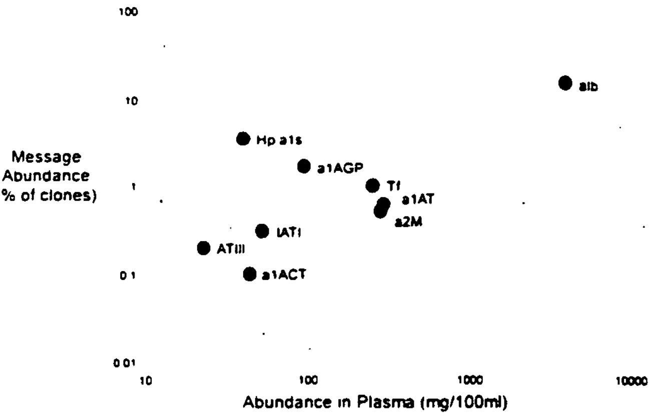


Figure 2. A log-log plot of data on mRNA abundance taken from Kawamoto *et al.* [12] *versus* average protein abundances in plasma taken from [21]. The protein abundance value for the haptoglobin α1s polypeptide has been corrected to reflect the fact that this subunit accounts for only 21% of the mass of the haptoglobin $\alpha_2\beta_2$ tetramer.

centage of total cDNA clones in the transcript images of two livers) indicates a modest correlation between the two (Fig. 1). The Pearson product-moment correlation coefficient obtained from the 19 pairs of measurements is 0.48. The abundance values obtained at the protein level spanned a 70-fold range, while the detectable mRNA abundances spanned a 16-fold range for these genes (although the latter value may reflect the limited number of clones sequenced). One particularly interesting subset of measurements concerns the β and γ actins. Here the mRNA abundances are, respectively, 0.189% and 0.215%, whereas the protein abundances are, respectively, 1.41% and 0.65% of the total. In this comparison, both sets of measurements are likely to be quite accurate, since numerous clones were detected for each of the two messages, and since the two proteins are so homologous, and have such close p*I*s, that they should bind CBB similarly. Nevertheless, the relative abundances at the RNA and protein levels are inverted (β actin is the more abundant protein, while γ actin has the more abundant message), and the mRNA:protein ratios for the two genes differ by more than a factor of two. Carbamyl phosphate synthase (CPS), the most abundant protein detected in liver over the p*I* range of conventional 2-D gels (pH 4–7), had a relative abundance of 2.83% (protein) and yet comprised only 0.139% of the total message (less than either actin). In this case, the mature protein is sequestered inside the mitochondrion, and therefore might be expected to show slow turnover and a consequent large disparity between mRNA and protein abundance.

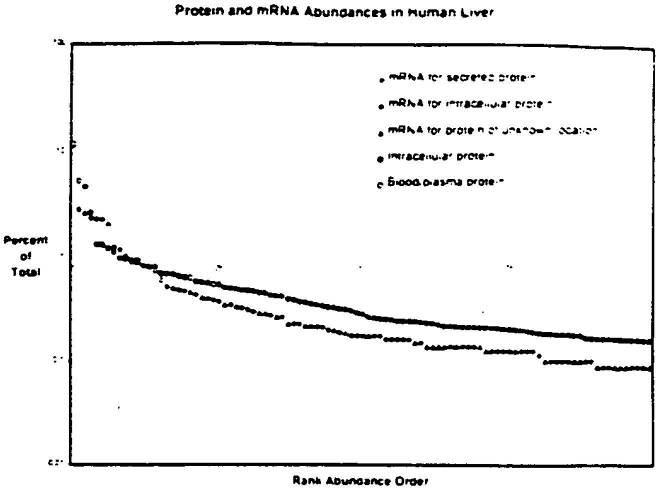


Figure 3. Relative abundance distributions of the top-ranked 100 mRNAs and proteins detected in human liver. The first (leftmost) molecule is the most abundant, followed by molecules of decreasing abundance through the 100th rank (at the right). Abundances of both mRNAs and proteins are plotted as a percentage of total detected molecules on a log scale. Message and protein points at the same rank are not, in general, products of the same gene.

A reexamination (Fig. 2) of the data of [12] on genes for plasma proteins, using estimates for corresponding protein abundances revised to account for the $\alpha_2\beta_2$ structure of haptoglobin, showed a higher correlation coefficient between mRNA and protein abundance (0.96). This value is probably exaggerated due to the large separation of the albumin values from the rest of the data: if albumin is omitted from the calculation, the correlation coefficient drops to –0.19. However, it is clear that the plasma proteins are represented by many more mRNA copies than major cellular proteins: albumin, for example, accounts for about 14% of the total number of clones examined [12], with a number of other plasma proteins accounting for more than 1% of the total each. By contrast, none of the cellular proteins chosen from the 2-D gel data accounted for much more than 0.1% of the mRNAs sequenced. To further pursue this observation, we compared the relative abundance distributions of the 100 top-ranked (most abundant) mRNAs and proteins in our data sets (Fig. 3). Forty-one of the top 100 mRNAs, and 29 of the top 50, coded for proteins known, or expected from sequence to be secreted from the liver, while none of the top 100 proteins appeared to be secretory forms of the human plasma proteins. The two most abundant proteins in these samples (hemoglobin β and albumin) as well as two of lower abundance (α_1 antiprotease and transferrin) were blood proteins that constitute contaminants of the liver in this context-proteins which would have been removed by perfusion.

4 Discussion

Despite extensive work on the regulation of many individual genes, little attention appears to have been paid to the global question of the relation between mRNA and corresponding protein abundance in eukaryotes. We have attempted to provide an initial estimate of the relationship of mRNA and corresponding cellular protein abundances through use of correspondences between two databases: the Molecular Anatomy™ (2-D gel) and LifeSeq® (Transcript Image) databases of human liver. Using a panel of 23 proteins identified on 2-D gels of human liver, we searched LifeSeq™ to determine the number of clones matching the corresponding gene sequence by BLAST. Matches were found for 19 proteins, and the correlation coefficient obtained over this set of data was 0.48. This number is intriguingly close to the middle position between a perfect correlation (1.0) and no correlation whatever (0.0). One simple interpretation of such a value is that the two major phases of gene expression regulation (transcription through message degradation on the one hand, and translation through protein degradation on the other) are of approximately equal importance in determining the net output of functional gene product (protein). Several issues may limit the quantitative accuracy of this result. First, the protein measurements rely on CBB binding to a series of different proteins. Although the measurements obtained show good (low) standard deviations across a set of six individual livers, it is well known that different proteins can bind CBB with different affinities. Thus the measurement scale for one protein may differ from another by up to approximately twofold. Since, however, these relative scale errors should be normally distributed, we expect them to have little effect on the overall correlation. Precision of the mRNA measurements is also limited, in this case because a limited number of clones was detected for the selected proteins. Five genes, for example, were represented by only one clone each among the 7925 clones sequenced from the respective cDNA tissue libraries. This low relative expression at the mRNA level is expected, since a majority of the high abundance mRNAs in liver code for plasma proteins. However, such small numbers of clones lead to potentially large quantitative errors because of sampling error. Here again, we believe these errors should be relatively random across the set of proteins chosen, and thus should not skew the result appreciably. A third potential difficulty is that the databases used for the protein and mRNA abundance estimates were prepared from different samples. In future, it will thus be of great interest to repeat the experiment using the same samples to examine both mRNA and protein abundances.

Despite these potential sources of error, at least one homologous pair of proteins (the β and γ actins) shows persuasive evidence of post-transcriptional regulation, with mRNA-to-protein ratios differing by more than a factor of two between the two genes. This is a particularly striking case since the two proteins are essentially indistinguishable in function (apart from affinity for MgADP; 22), have very similar sequences, and are produced in a constant ratio (approximately 2:1 in males) in virtually all cell types. One possible alternative explanation could be a sex difference in liver expression of γ actin, as is seen in rodents [23] where γ actin protein expression averages almost twice as high in females as males. This seems unlikely since 64% of clones in the RTI data were from male liver, and all the 2-D data was from male livers.

An analogous set of data for plasma proteins secreted by the liver has been published by Kawamoto *et al.* [12] and we have reanalyzed their values to see whether a similar mRNA-to-protein relationship holds. It appears, based on nine plasma proteins, that a higher correlation coefficient applies: 0.96. This result is less convincing, however, because one gene product (albumin) is well-separated from the cluster of the remaining eight, and thus exercises a disproportionate influence on the correlation coefficient. In fact, if albumin is omitted from the calculation, the correlation coefficient is reduced to −0.19, which suggests a very poor correlation.

What is perhaps more striking is the relatively much higher abundance of the plasma protein mRNAs as compared to major cellular proteins such as carbamyl phosphate synthase, the actins, or cytochrome b5. Mid-abundance plasma proteins were represented by mRNAs having approximately 100-fold higher relative abundance than mid-abundance cellular proteins. This result is verified by a direct comparison of the relative abundance distributions of the 100 top-ranked mRNAs and proteins in our data sets (which are, in general, different sets of genes). Twenty-nine of the top 50 messages are secreted products, while none of the top 50 proteins appear to be the pro- form of a secreted molecule. Such a conclusion is not surprising, since the liver is responsible for generating high protein concentrations in the relatively large plasma compartment of the body, but does so by means of closely coupled synthesis and secretion with little accumulation of precursor proteins in process. This points to a potentially significant difference in the pictures obtained from mRNA and protein abundance databases. Major secreted proteins appear to have much more abundant mRNAs than many important cellular proteins, and hence mRNA abundance databases that concentrate on a small number of the highest abundance messages may be biased towards secreted proteins over cellular molecules. This represents an advantage of the mRNA approach relative to protein databases in the search for novel cytokines and other secreted proteins, but a disadvantage in the characterization of cellular metabolic and control processes. Additionally, it suggests that mRNAs for secreted proteins may have, on the whole, shorter half-lives than mRNAs for cellular enzymes, the latter being more frequently regulated at the translational level.

We also found important differences in the overall shapes of the relative abundance distributions of the 100 top-ranked mRNAs and proteins. While both distributions contain a few very high abundance molecules (in the 3–10% range) they appear to diverge significantly below the 15th most abundant gene product, with proteins 16–100 accounting for roughly twice as high a relative abundance as the 16th–100th mRNAs. Not all proteins are represented on the 2-D gels used here (which fail to resolve proteins with p*I* >7), but the estimated 40% of proteins thus excluded would not affect the shape of the distribution over positions 50–100 significantly if they have an abundance distribution similar to the p*I* 4–7 proteins (based on a simulation using the data shown). The mRNA abundance distribution covers all cloned messages (not a subset of genes), and for abundant mRNAs it should be complete as it stands. Altogether, the top 100 mRNAs comprise 51.3% of the total clones, while the top 100 proteins comprise 63.1% of the total protein detected. Hence it appears likely that the distribution of protein abundances is significantly different from that of mRNAs, showing a more gradual fall-off in the region examined, and that techniques able to detect down to a specified percent abundance threshold would reveal more proteins at a given threshold than mRNAs. As the protein and nucleic acid databases expand, we anticipate the possibility of generating successively more robust estimates of the global relationship between mRNA and protein abundance, and thus a better understanding of multi-level gene expression control in complex organisms such as man.

Human liver samples analyzed by 2-D electrophoresis were kindly provided by the National Biomonitoring Specimen Bank at the US National Institute of Standards and Technology under the direction of Dr. Stephen Wise.

Received November 27, 1996

5 References

[1] Okubo, K., Hori, N., Matoba, R., Niiyama, T., Matsubara, K. A., *Nat. Genet.* 1992, *2*, 173–179.

[2] Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K. W., *Science* 1995, *270*, 484–487.

[3] Liang, P., Pardee, A. B., *Curr. Opin. Immuno.* 1995, *7*, 274–28.

[4] Augenlicht, L. H., Wahrman, M. Z., Halsey, H., Anderson, L., Taylor, J., Lipkin, M., *Cancer Res.* 1987, *47*, 6017–6021.

[5] Fodor, S. P., Rava, R. P., Huang, X. C., Pease, A. C., Holmes, C. P., Adams, C. L., *Nature* 1993, *364*, 555–556.

[6] Schena, M., Shalon, D., Davis, R. W., Brown, P. O., *Science* 1995, *270*, 467–470.

[7] O'Farrell, P. H., *J. Biol. Chem.* 1975, *250*, 4007–4021.

[8] Garrels, J. I., Futcher, B., Kobayashi, R., Latter, G. I., Schwender, B., Volpe, T., Warner, J. R., McLaughlin, C. S., *Electrophoresis* 1994, *15*, 1466–1486.

[9] Celis, J. E., Rasmussen, H. H., Olsen, E., Madsen, P., Leffers, H., Honoré, B., Dejgaard, K., Gromov, P., Vorum, H., Vassilev, A., Baskin, Y., Liu, X., Celis, A., Basse, B., Lauridsen, J. B., Ratz, G. P., Andersen, A. H., Walbum, E., Kjærgaard, A., Andersen, S., Puype, M., Van Damme, J., Vanderkerckove, J., *Electrophoresis* 1994, *15*, 1349–1458.

[10] Hochstrasser, D. F., Frutiger, S., Paquet, N., Bairoch, A., Ravier, F., Pasquali, C., Sanchez, J. C., Tissot, J. D., Bjellqvist, B., Vargas, R., Appel, R. D., Hughes, G. J., *Electrophoresis* 1992, *13*, 992–1001.

[11] Anderson, N. L., Esquer-Blasco, R., Hofmann, J. P., Meheus, L., Raymackers, J., Steiner, S., Witzmann, F., Anderson, N. G., *Electrophoresis* 1995, *16*, 1977–1981.

[12] Kawamoto, S., Matsumoto, Y., Mizuno, K., Okubo, K., Matsubara, K., *Gene* 1996, *174*, 151–158.

[13] Anderson, N. L., Anderson, N. G., *Anal. Biochem.* 1978, *85*, 341–354.

[14] Anderson, N. G., Anderson, N. L., *Anal. Biochem.* 1978, *85*, 331–340.

[15] Anderson, N. L., Esquer-Blasco, R., Hofmann, J.-P., Anderson, N. G., *Electrophoresis* 1991, *12*, 907–930.

[16] Anderson, N. L., Large Scale Biology Press, Washington, DC 1991, ISBN 0-945532-01-6, 200 pp., and http://www.lsbc.com.

[17] Hughes, G. J., Frutiger, S., Paquet, N., Pasquali, C., Sanchez, J.-C., Tissot, J.-D., Bairoch, A., Appel, R., Hochstrasser, D., *Electrophoresis* 1993, *14*, 1216–1222.

[18] Chirgwin, J., Przybla, A., MacDonald, R., Rutter, W., *Biochemistry* 1979, *18*, 5294–5299.

[19] Gubler, U., Hoffman, B. J., *Gene* 1983, *25*, 263–269.

[20] Sanger, F., Nicklen, S., Coulson, A. R., *Proc. Nat. Acad. Sci. USA* 1977, *74*, 5463–5469.

[21] Putnam, F. W., (Ed.), *The plasma proteins*, Academic Press, New York 1975, pp. 26–29.

[22] Anderson, N. L., *Biochem. Biophys. Res. Comm.* 1979, *89*, 486–490.

[23] Anderson, N. L., Giere, F. A., Nance, S. L., Gemmell, M. A., Tollaksen, S. L., Anderson, N. G., Galteau, M.-M., Siest, G. (Eds.), *Progrès Récents en Electrophorèse Bidimensionelle*, Presses Universitaires de Nancy, Nancy 1986, pp. 253–260.

B

N. Leigh Anderson[1]
Ricardo Esquer-Blasco[1]
Jean-Paul Hofmann[1]
Lydie Meheus[2]
Jos Raymackers[2]
Sandra Steiner[3]
Frank Witzmann[4]
Norman G. Anderson[1]

[1]Large Scale Biology Corporation, Rockville, MD
[2]Innogenetics NV, Ghent
[3]Sandoz Pharma Ltd, Drug Safety Assessment, Toxicology, Basel
[4]Molecular Anatomy Laboratory, Indiana University Purdue University Columbus, Columbus, IN

An updated two-dimensional gel database of rat liver proteins useful in gene regulation and drug effect studies

We have improved upon the reference two-dimensional (2-D) electrophoretic map of rat liver proteins originally published in 1991 (N. L. Anderson *et al.*, *Electrophoresis* 1991, *12*, 907–930). A total of 53 proteins (102 spots) are now identified, many by microsequencing. In most cases, spots cut from wet, Coomassie Blue stained 2-D gels were submitted to internal tryptic digestion [2], and individual peptides, separated by high-performance liquid chromatography (HPLC), were sequenced using a Perkin-Elmer 477A sequenator. Additional spots were identified using specific antibodies.

Figure 1 shows the current annotated 2-D map of F344 rat liver, analyzed using the Iso-DALT system (20 × 25 cm gels) and BDH 4–8 carrier ampholytes. Both the map itself and the master spot number system remain the same as shown in the original publication. Table 1 lists the important features of each identification shown, including the gel position, p*I*, and M_r for the most abundant or most basic form of each protein. Using this extended base of identified spots, a series of four improved calibration functions has been derived for the p*I* and SDS-M_r axes (the first two of which are shown in Fig. 2A and B). Both forward and reverse functions are derived, so that one can compute the physical properties of a spot with a given gel location, or inversely compute the gel position expected for a protein having given physical properties:

$$Y_{\text{RATLIVER}} = f_{\text{M}-\text{RATLIVER Y}}(M_{r,\text{SEQUENCE-DERIVED}}) \qquad (1)$$

$$X_{\text{RATLIVER}} = f_{\text{pI}-\text{RATLIVER X}}(\text{p}I_{\text{SEQUENCE-DERIVED}}) \qquad (2)$$

$$M_{r,\text{GEL-DERIVED}} = f_{\text{RATLIVER Y}-\text{M}_r}(Y_{\text{RATLIVER}}) \qquad (3)$$

$$\text{p}I_{\text{GEL-DERIVED}} = f_{\text{RATLIVER X}\to\text{pI}}(X_{\text{RATLIVER}}) \qquad (4)$$

A spreadsheet program (in Microsoft Excel) was developed to facilitate flexible computation of p*I*'s from amino acid sequence data, and the results were entered into a relational database (Microsoft Access). A table of spot positions and sequence-derived pI's and M_r's was fitted with a large series of analytic equations using Tablecurve (Jandel Scientific), and the four conversion Eqs. (1)–(4), relating computed p*I* and gel *X* coordinate, or computed molecular weight and gel *Y* coordinate, were selected, based on criteria of simplicity, goodness of fit and favorable asymptotic behavior. Table 2 lists the equations and coefficients. Application of Eqs. (3) and (4) to a spot's *X* and *Y* coordinates, given in [1], produce improved M_r estimates, and allow computation of p*I* directly in pH units, instead of in terms of positions relative to creatine phosphokinase (CPK) charge standards. The inverse Eqs. (1) and (2) were used to compute the gel positions of a series of p*I* and M_r tick marks. These tick marks were plotted with SigmaPlot (Jandel), together with fiducial marks locating several prominent spots, and the resulting graphic was aligned over the synthetic gel image (computed by Kepler from the master gel pattern) using Freelance (Lotus Development). Maps were printed as Postscript output from Freelance, either in black and white (as shown here) or in color, where label color indicates subcellular location (available from the first author upon request). We have also used the rat liver 2-D pattern as presented here to calibrate the patterns of other samples. Using mixtures of rat liver and mouse liver samples, for example, we made composite 2-D patterns that allow use of the rat pattern to standardize both axes of the mouse pattern. This was accomplished by deriving transformations relating the rat and mouse *X*, and separately the rat and mouse *Y*, axes (Table 2, lower half; Fig. 2C and D) based on a series of spots that coelectrophorese in these closely related species. These functions were then applied to derive equations relating the mouse liver *X* and *Y* to p*I* and SDS-M_r (Eqs. 5 and 6 below). The resulting standardized 2-D pattern for B6C3F1 mouse liver is shown in Fig. 3.

$$M_{r,\text{MOUSELIVER}} = f_{\text{RATLIVER Y}-\text{M}_r}(f_{\text{MOUSELIVER Y}-\text{RATLIVER Y}}(Y_{\text{MOUSELIVER}})) \qquad (5)$$

$$\text{p}I_{\text{MOUSELIVER}} = f_{\text{RATLIVER X}\to\text{pI}}(f_{\text{MOUSELIVER X}-\text{RATLIVER X}}(X_{\text{MOUSE LIVER}})) \qquad (6)$$

A slightly more complex approach can be used to standardize samples that have few or no spots co-electrophoresing with rat liver proteins. In this case, a 2-D gel is prepared with a mixture of the two samples, and four functions (forward and backward, each for *X* and *Y*) are derived relating each sample's own master pattern to the composite. The required functions are then applied in a nested fashion to yield the desired result (using rat plasma as an example):

$$M_{r,\text{RATPLASMA}} = f_{\text{RATLIVER Y}-\text{M}_r}(f_{\text{RATPLASMA+LIVER Y}-\text{RATLIVER Y}}(f_{\text{RATPLASMA Y}-\text{RATPLASMA+LIVER Y}}(Y_{\text{RATPLASMA}}))) \qquad (7)$$

Correspondence: Dr. Leigh Anderson, Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338 USA (Tel: +301-424-5989; Fax: +301-762-4892; email: leigh@lsbc.com)

Keywords: Two-dimensional polyacrylamide gel electrophoresis / Liver / Map / Identification / Calibration

0173-0835/95/1010-1977 $5.00+.25/0

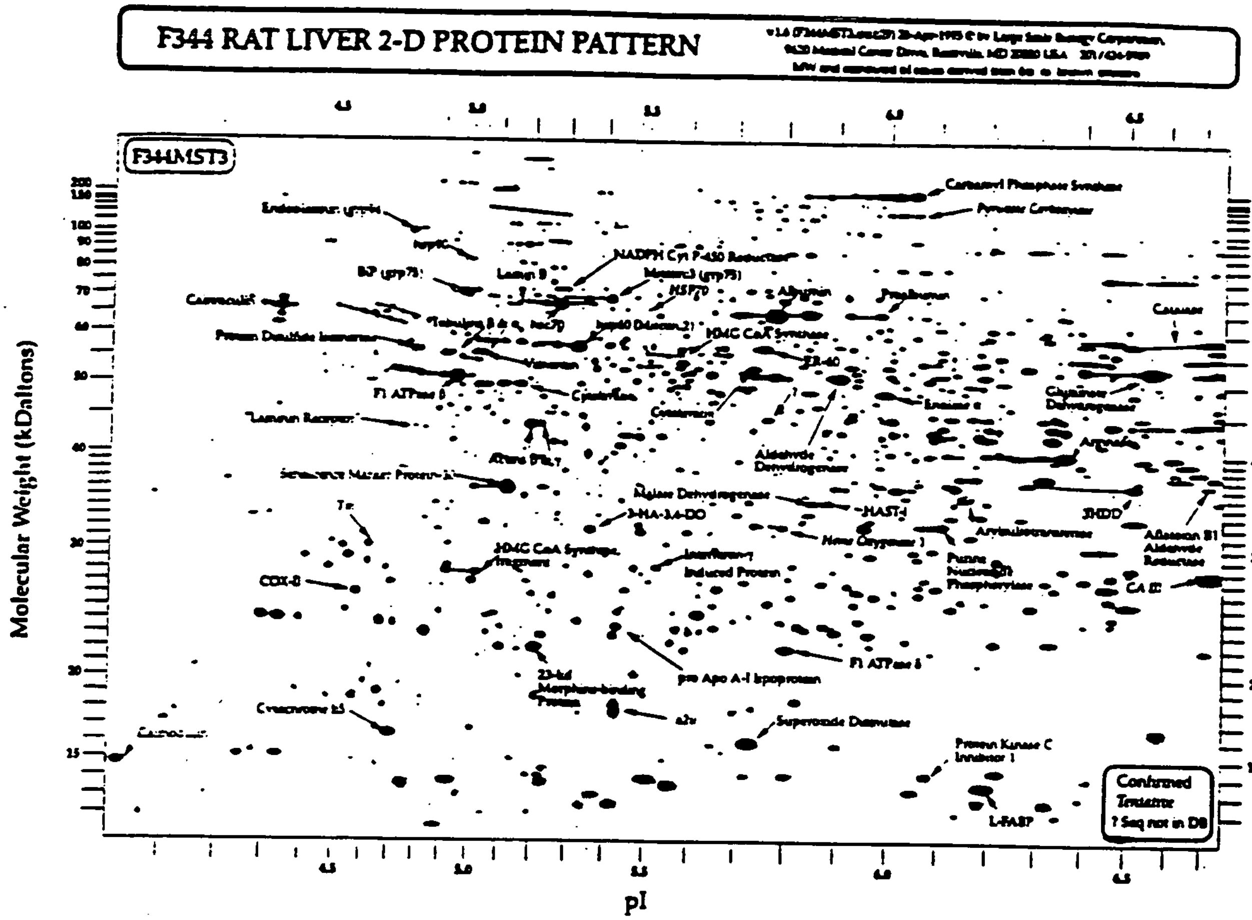


Figure 1. Master 2-D gel pattern of Fischer 344 rat liver proteins, annotated with 53 protein identifications and computed p*I* and M_r axes. Tentative identifications are in italic type.

Table 1. Proteins identified in the 2-D pattern of F344 rat liver

MSN[a)]	Protein ID[b)]	Protein name	Identification comments	Gel X[c)]	Experimental p*I*[d)]	Gel Y[c)]	Experimental M_r[d)]
126	HADO-HUMAN[e)]	3-HA-3,4-DO: 3-hydroxy-anthranilate-3,4-dioxygenase	Internal sequence	871.95	5.36	921.35	30 207
137, 159, 288, 258	DIDH_RAT	3HDD: 3-hydroxysteroid dihydrodiol reductase	Ab (T.M. Penning) and pure protein	1857.52	6.51	822.52	34 406
173	MUP_RAT	a_2u globulin	Presence in liver microsome lumen, abundance in kidney, p*I*, M_r	919.16	5.43	1313.81	19 549
38	ACTB_HUMAN	Actin β	Analogy with other mammalian patterns (*e.g.* human) through coelectrophoresis	763.40	5.19	693.64	41 586
68	ACTG_HUMAN	Actin γ	Analogy with other mammalian patterns (*e.g.* human) through coelectrophoresis	779.42	5.21	692.26	41 677
693	AFAR_RAT	Aflatoxin B1 aldehyde reductase	Internal sequence	1993.32	6.72	818.60	34 593
28, 21, 33	ALBU_RAT	Albumin	Coelectrophoresis with principal plasma protein	1262.81	5.86	445.64	66 354
43	DHAM_RAT	Aldehyde dehydrogenase	N-Terminal sequence and AAA	1317.72	5.91	589.03	49 602
96	ARGI_RAT	Arginase	Internal sequence	1730.72	6.34	756.02	37 819
117	SUAR_RAT	Arylsulfotransferase	Internal sequence	1547.96	6.14	849.08	33 186
1163, 1161, 1162, 20	GR78_RAT	BIP (GRP-78)	Ab (F. Witzmann)	665.33	5.01	397.39	74 564
185	CAH3_RAT	CA-III	Uncertain; by comparison with mouse	1996.60	6.72	1017.02	26 887
123	CALM_HUMAN	Calmodulin	Analogy with human cellular patterns through coelectrophoresis	23.05	4.03	1433.25	17 419
3, 201, 48, 39, 22, 24	CRTC_RAT	Calreticulin	Ab (Lance Pohl)	310.59	4.34	433.80	68 206

Table 1. continued

MSN[a]	Protein IDb)	Protein name	Identification comments	Gel X[c]	Experimental pI[c]	Gel Y[c]	Experimental M_r[d]
1184, 1186, 114, 174, 118 5, 167, 157	CPSM_RAT	Carbamyl phosphate synthase	2-D of pure protein; comfirmed by *N*-terminal sequence and AAA	1453.56	6.05	181.64	160 640
54, 61	CATA_RAT	Catalase	Internal sequence	2000.81	6.73	499.64	58 968
136	COX2_RAT	COX-II	Ab (J. W. Taanman), confirmed by internal sequence	452.57	4.61	1062.67	25 504
87	CYB5_RAT	Cytochrome B5	2-D of pure protein; Ab; confirmed by AAA	515.68	4.73	1370.55	18 493
41	CK-RAT[e]	Cytokeratin	Location in cytoskeletal fraction	1165.12	5.75	569.09	51 448
29	CK-RAT[e]	Cytokeratin	Location in cytoskeletal fraction	743.11	5.15	605.23	48 187
5, 11	ENPL-RAT[e]	Endoplasmin	Ab (F. Witzmann)	567.73	4.83	263.37	112 194
60	ENOA_RAT	Enolase A	Internal sequence and AAA	1399.78	6.00	623.54	46 674
27	ER60_RAT	ER-60	*N*-Terminal sequence (R. M. Van Frank)	1184.20	5.77	523.51	56.169
17	ATPB_RAT	F1 ATPase β	*N*-Terminal sequence and AAA	629.06	4.95	588.83	49 620
196	ATP7_RAT	F1 ATPase δ	Internal sequence	1227.24	5.82	1184.65	22 310
79	F16P_RAT	Fructose-1.6-bis-phosphatase	Uncertain; by comparison with ID in Garrison and Wager (JBC 257:13135-13143)	924.54	5.44	737.77	38 858
62, 78	DHE3_RAT	Glutamate dehydrogenase	*N*-Terminal sequence and internal sequence	1887.39	6.55	566.92	51 655
125	HAST-RAT[e]	HAST-I: N-hydroxyaryl-amine sulfotransferase	Internal sequence	1297.94	5.89	861.55	32 638
307	HO1_RAT	Heme oxygenase 1	Uncertain; available data from internal sequence	1219.39	5.81	915.71	30 423
413, 1250, 933	HMCS_RAT	HMG CoA synthase, cytosolic	Ab (J. Germershausen)	1033.48	5.59	538.13	54 571
133, 144, 235	HMCS_RAT	HMG CoA synthase, mitochondrial (frag)	Ab (J. Germershausen), *N*-terminal sequence (Steiner/Lottspeich)	666.40	5.02	1019.42	26 811
8, 23, 1307	HS7C_RAT	HSC-70	Positional homology (with human, *etc.*) through coelectrophoresis	811.87	5.27	425.76	69 521
15, 25, 110	P60_RAT	HSP-60	Ab (F. Witzman); confirmed by *N*-terminal sequence and AAA	845.09	5.32	520.03	56 561
971	HS70-RAT[e]	HSP-70	Ab (F. Witzman)	976.11	5.51	437.14	67 674
1216, 1215, 90	HS90-RAT[e]	HSP-90	Ab (F. Witzman)	659.86	5.00	329	90 107
256	INGI-HUMAN	Interferon-γ induced protein	Internal sequence	993.85	5.54	1006.04	27 237
415, 734	LAMB-RAT[e]	Lamin B	Positional homology with human through coelectrophoresis, nuclear location	737.10	5.14	425.19	69 615
80	LAMR-RAT[e]	"Laminin receptor"	Internal sequence	534.02	4.77	697.62	41 327
227	FABL_RAT	L-FABP (liver fatty acid binding protein)	Ab (N. M. Bass)	1586.09	6.18	1483.43	16 622
134	MDHC_MOUSE	Malate dehydrogenase	Internal sequence	1270.85	5.86	861.96	32 620
18, 35, 226	GR75-RAT[e]	Mitcon:3; grp75	Positional homology with human through coelectrophoresis	905.67	5.41	413.67	71 589
175, 251	NCPR_RAT	NADPH P450 reductase	2-D of pure protein	824.69	5.29	393.21	75 366
1168, 1170, 1171	PDI_RAT	PDI: Protein disulfide isomerase	*N*-Terminal sequence (R. M. van Frank), Ab	564.30	4.83	528.47	55 618
47, 93	ALBU_RAT	Pro-Albumin	Microsomal lumen location, *pI*, M_r relative to albumin	1391.03	5.99	446.68	66 195
236	APA1_RAT	Pro-APO A-I lipoprotein	Coelectrophoresis with plasma protein	920.41	5.43	1137.51	23 467
320	IPK1_BOVIN	Protein kinase C inhibitor 1	Internal sequence; homology with bovine protein	1480.01	6.08	1458.81	17 007
152	PNPH_MOUSE	Purine nucleoside phosphorylase	Internal sequence	1507.19	6.10	911.16	30 599
1179, 1180, 1181, 1182, 1183	PYVC-RAT[e]	Pyruvate carboxylase	Tentative; 2-D of pure protein (J. G. Henslee, *JBC*, 1979); reported in *Biochim. Biophys. Acta* *1022*, 115–125	1485.10	6.08	223.52	131 589
55, 103	SM30_RAT	SMP-30: Senescence marker protein-30	Internal sequence	721.71	5.11	830.10	34 051
135	SODC_RAT	Superoxide dismutase	AAA; comfirmed by internal sequence (R. M. Van Frank)	1161.24	5.74	1388.68	18 173
172	TPM-RAT[e]	Tm: tropomyosin	Location in cytoskeleton, 2-D position relative to human, Ab	476.24	4.66	957.86	28 865
277, 56	TBA1_RAT	Tubulin α	Positional homology with human through coelectrophoresis, cytoskeletal location	688.22	5.06	537.67	54 620
50, 1225	TBB1_RAT	Tubulin β	Positional homology with human through coelectrophoresis, cytoskeletal location	621.29	4.93	535.48	54 855
1224	VIME_RAT	Vimentin	Positonal homology with human through coelectrophoresis, cytoskeletal location	673.00	5.03	539.50	54 426

Table 1. continued

MSN[a]	Protein ID[b]	Protein name	Identification comments	Gel X[c]	Experimental pI[d]	Gel Y[c]	Experimental M_r[d]
113	Unknown	?: not in sequence databases	Internal sequence	1191.28	5.78	680.42	42 469
104	BBPL_RAT	23 kDa morphine-binding protein	Internal sequence	773.31	5.20	1182.41	22 363

a) Master spot number (MSN) from [1]
b) SwissPROT identifier
c) Coordinates of the most basic or most abundant assigned spot on the F344 master gel pattern
d) pI and M_r of the most basic or most abundant assigned spot, derived from the calibration functions included here
e) SwissPROT style proposed identifier
Abbreviations: AAA, amino acid analysis; Ab, antibody

Table 2. Equations and coefficients

Function	Equation (f)	r2	a	b	c	d	e
Rat gel Y = f(computed M_r)	$y = a + b\exp(-x/c)$	0.988181021	178.74803	1967.7892	32363.958		
Rat gel X = f(computed pI)	$y = a + bx + cx/\ln x + d/x + e/x^{1.5}$	0.99247216	−8685665.5	−904497.94	3856926.1	18276844	−27154534
Computed M_r = f(rat gel Y)	$y = a + bxc$	0.9960177	−8464.5809	19095881	−0.9086255		
Computed pI = f(rat gel X)	$y = a + bx + cx^2 + dx^2 \ln x + ex^3$	0.99176499	4.044686	−0.00114238	0.0000323	−0.00000455	0.0000000176
Mouse gel Y = f(rat gel Y)	$y = a + bx + cx^{1.5} + dx^{0.5} \ln x + ex/\ln x$	0.99951069	11861.44	678.91666	−0.78964914	1567.5639	−6953.9592
Mouse gel X = f(rat gel X)	$y = a + bx^2 \ln x + cx^{2.5} + dx^3$	0.99926349	58.935923	0.00091353	−0.000213688	0.00000159	
Rat gel Y = f(mouse gel Y)	$y = a + bx^2 \ln x + cx^{2.5} + dx^3$	0.99950032	69.740526	0.00050772	−0.000130392	0.00000116	
Rat gel X = f(mouse gel X)	$y = a + bx + cx^2 \ln x + dx^{2.5} + ex^3$	0.9992832	−198.07189	2.0899063	−0.000671191	0.000145189	−0.000000986

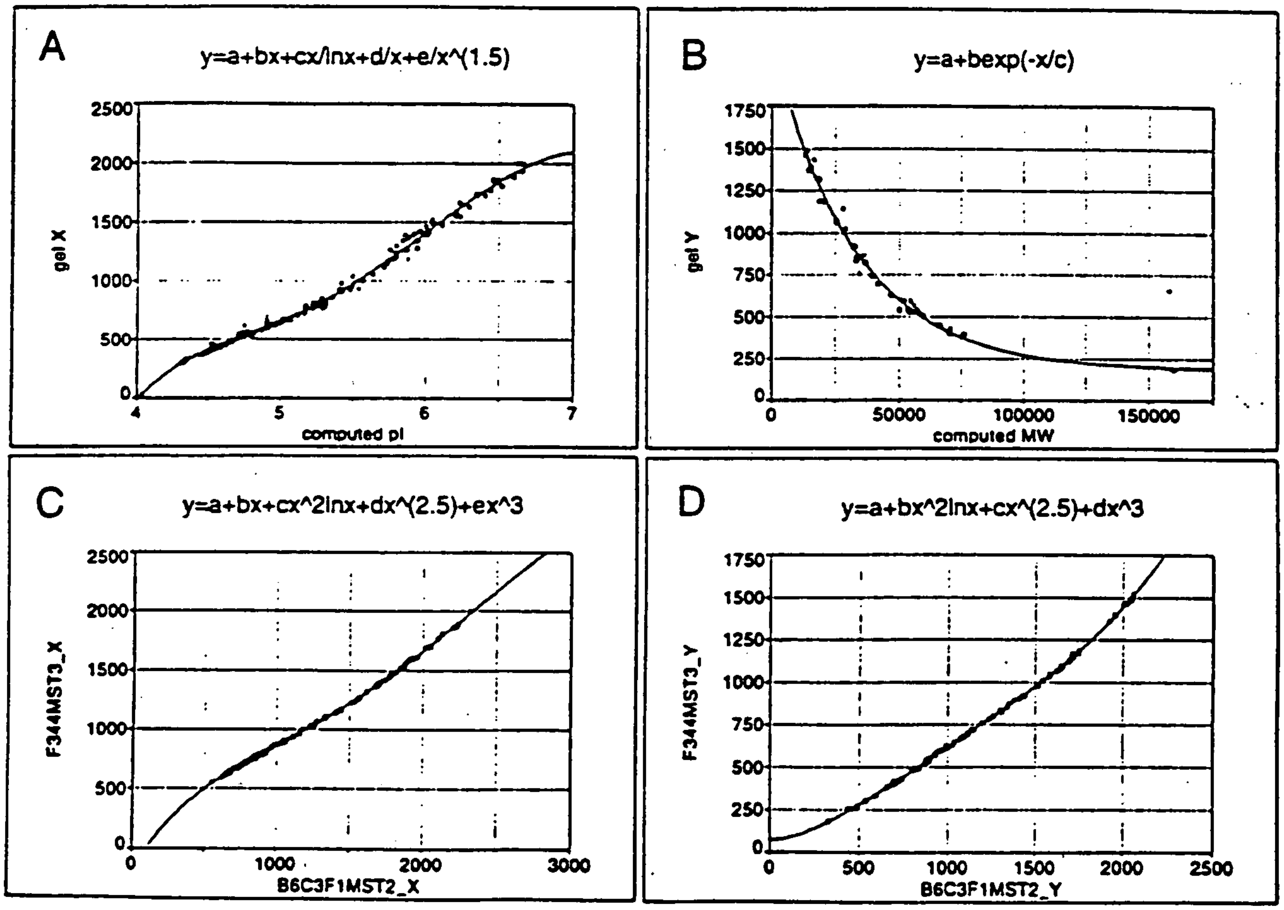


Figure 2. Plots showing fits of selected equations (continuous curves) to data on identified proteins (square symbols). (A) pI computed from sequence data *versus* gel X position for identified spots in F344 rat liver; (B) M_r computed from sequence data *versus* gel Y position for identified spots in F344 rat liver; (C) gel X position for spots in B6C3F1 mouse liver *versus* X position in F3443 rat liver, for coelectrophoresing spots; (D) gel Y position for spots in B6C3F1 mouse liver *versus* Y position in F3443 rat liver, for coelectrophoresing spots. In each case, inverse equations were also computed (Table 2).

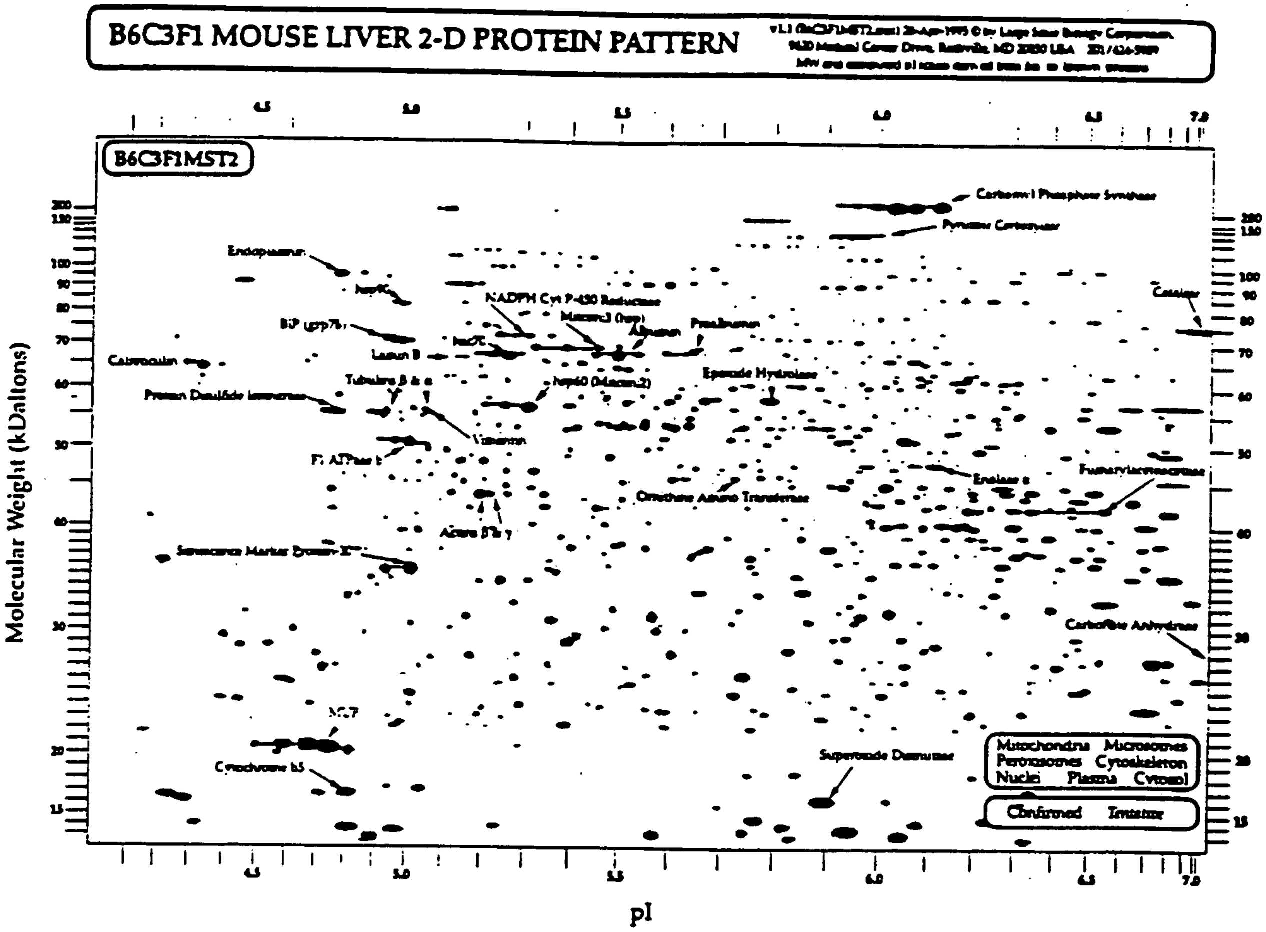


Figure 3. Master 2-D gel pattern for B6C3F1 mouse liver, standardized using the F344 rat liver pattern identifications, according to the method described in the text. Twenty-nine proteins are identified.

$$pI_{RATPLASMA} = f_{RATLIVER\ X\to pI}(f_{RATPLASMA\to LIVER\ X\to RATLIVER\ X}(f_{RATPLASMA\ X\to RATPLASMA\to LIVER\ X}(X_{RAT\ PLASMA}))) \qquad (8)$$

This unified approach, in which one well-populated 2-D pattern is used to standardize a family of other patterns, has the additional advantage that the resulting p*I* and *M*, scales are directly compatible. Hence one can compare the relative p*I*'s of mouse and rat versions of a sequenced protein in a consistent p*I* measurement system, and select likely inter-species analogs based on positional relationships on common scales. Adoption of immobilized pH gradient (IPG) technology [4–7] will result in substantial improvements in p*I* positional reproducibility for standard 2-D maps such as those presented here; however, we believe that our approach will continue to be useful in establishing the empirical pH gradient actually achieved by such gels under given experimental conditions (temperature, urea concentration, *etc.*), in relating patterns run on different IPG ranges and using different lots of IPG gels (between which some variation will persist). Development of rodent organ maps is a continuing effort in our laboratories [8–10], and results in regular additions of identified proteins. Those who wish to receive current rodent liver maps, with color annotations, should send a stamped self-addressed envelope to the first author.

We would like to thank the individuals who provided antibodies mentioned in Table 1, and R. M. van Frank for unpublished sequenced data.

Received May 31, 1995

References

[1] Anderson, N. L., Esquer-Blasco, R., Hofmann, J.-P., Anderson, N. G., *Electrophoresis* 1991, *12*, 907–930.

[2] Rosenfeld, J., Capdevielle, J., Guillemot, J. C., Ferrara, P., *Anal. Biochem.* 1992, *203*, 173–179.

[3] Witzmann, F., Clack, J., Fultz, C., Jarnot, B., *Electrophoresis* 1995, *16*, 451–459.

[4] Rosengren, A. E., Bjellqvist, B., Gasparic, V., *US Patent* 4130470, December 1978.

[5] Gianazza, E., Artoni, G., Righetti, P. G., *Electrophoresis* 1983, *4*, 321–326.

[6] Görg, A., Postel, W., Günther, S., Weser, J., *Electrophoresis* 1985, *6*, 599–604.

[7] Gianazza, E., Astrua-Testori, S., Giacon, P., Righetti, P. G., *Electrophoresis* 1985, *6*, 332–339.

[8] Myers, T. G., Dietz, E. C., Anderson, N. L., Khairallah, E. A., Cohen, S. D., Nelson, S. D., *Chem. Res. Toxicol.* 1995, *8*, 403–413.

[9] Cunningham, M. L., Pippin, L. L., Anderson, N. L., Wenk, M. L., *Toxicol. Appl. Pharmacol.* 1995, *131*, 216–223.

[10] Anderson, N. L., Copple, D. C., Bendele, R. A., Probst, G. S., Richardson, F. C., *Fundam. Appl. Toxicol.* 1992, *18*, 570–580.

C

2

Progress with Proteome Projects: Why all Proteins Expressed by a Genome Should be Identified and How To Do It

MARC R. WILKINS[1], JEAN-CHARLES SANCHEZ[3], ANDREW. A. GOOLEY[1], RON D. APPEL[3], IAN HUMPHERY-SMITH[2], DENIS F. HOCHSTRASSER[3] AND KEITH L. WILLIAMS[1,*]

[1] Macquarie University Centre for Analytical Biotechnology, Macquarie University, Sydney, NSW 2109, Australia; [2] Department of Microbiology, University of Sydney, NSW, 2006, Australia and [3] Central Clinical Chemistry Laboratory and Medical Computing Centre of the University of Geneva, CH 1211 Geneva 14, Switzerland

Introduction

The advent of large genome sequencing projects has changed the scale of biology. Over a relatively short period of time, we have witnessed the elucidation of the complete nucleotide sequence for bacteriophage λ (Sanger *et al.*, 1982), the sequence of an eukaryotic chromosome (Oliver *et al.*, 1992), and in the near future will see the definition of all open reading frames of some simple organisms, including *Mycoplasma pneumoniae*, *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Arabidopsis thaliana*. Nevertheless, genome sequencing projects are not an end in themsleves. In fact, they only represent a starting point to understanding the function of an organism. A great challenge that biologists now face is how the co-expression of thousands of genes can best be examined under physiological and pathophysiological conditions, and how these patterns of expression define an organism.

There are two approaches that can be used to examine gene expression on a large scale. One uses nucleic acid-based technology, the other protein-based technology. The most promising nucleic-acid based technology is differential display of mRNA (Liang and Pardee, 1992; Bauer *et al.*, 1993), which uses polymerase chain reaction with arbitrary primers to generate thousands of cDNA species, each which correspond to an expressed gene or part of a gene. However, it is currently unclear if this technique can be developed to reliably assay the expression of thousands of genes or

[*] Corresponding Author

Biotechnology and Genetic Engineering Reviews – Vol. 13, December 1995
0264–8725/94/12/1–38 $20.00 + $0.00 © Intercept Ltd, P.O. Box 716, Andover, Hampshire SP10 1YG, UK

identify all cDNA species, and the approach does not easily allow a systematic screening. Analysis of gene expression by the study of proteins present in a cell or tissue presents a favorable alternative. This can be achieved by use of two-dimensional (2-D) gel electrophoresis, quantitative computer image analysis, and protein identification techniques to create 'reference maps' of all detectable proteins. Such reference maps establish patterns of normal and abnormal gene expression in the organism, and allow the examination of some post-translational protein modifications which are functionally important for many proteins. It is possible to screen proteins systematically from reference maps to establish their identities.

To define protein-based gene expression analysis, the concept of the 'proteome' was recently proposed (Wilkins *et al.*, 1995; Wasinger *et al.*, 1995). A proteome is the entire PROTein complement expressed by a genOME, or by a cell or tissue type. The concept of the proteome has some differences from that of the genome, as while there is only one definitive genome of an organism, the proteome is an entity which can change under different conditions, and can be dissimilar in different tissues of a single organism. A proteome nevertheless remains a direct product of a genome. Interestingly, the number of proteins in a proteome can exceed the number of genes present, as protein products expressed by alternative gene splicing or with different post-translational modifications are observed as separate molecules on a 2-D gel. As an extrapolation of the concept of the 'genome project', a 'proteome project' is research which seeks to identify and characterise the proteins present in a cell or tissue and define their patterns of expression.

Proteome projects present challenges of a similar magnitude to that of genome projects. Technically, the 2-D gel electrophoresis must be reproducible and of high resolution, allowing the separation and detection of the thousands of proteins in a cell. Low copy number proteins should be detectable. There should be computer gel image analysis systems that can qualitatively and quantitatively catalog the electrophoretically separated proteins, to form reference maps. A range of rapid and reliable techniques must be available for the identification and characterisation of proteins. As a consequence of a proteome project, protein databases must be assembled that contain reference information about proteins; such databases must be linked to genomic databases and protein reference maps. Databases should be widely accessible and easy to use.

Recently, there have been many changes in the techniques and resources available for the analysis of proteomes. It is the aim of this chapter to discuss the status of the areas outlined above, and to review briefly the progress of some current proteome projects.

Two-dimensional electrophoresis of proteomes

Two dimensional (2-D) gel electrophoresis involves the separation of proteins by their isoelectric point in the first dimension, then separation according to molecular weight by sodium dodecyl sulfate electrophoresis in the second dimension. Since first described (Klose, 1975; O'Farrell, 1975; Scheele, 1975), it has become the method of choice for the separation of complex mixtures of proteins, albeit with many modifications to the original techniques. 2-D electrophoresis forms the basis of proteome projects through separating proteins by their size and charge (Hochstrasser *et al.*,

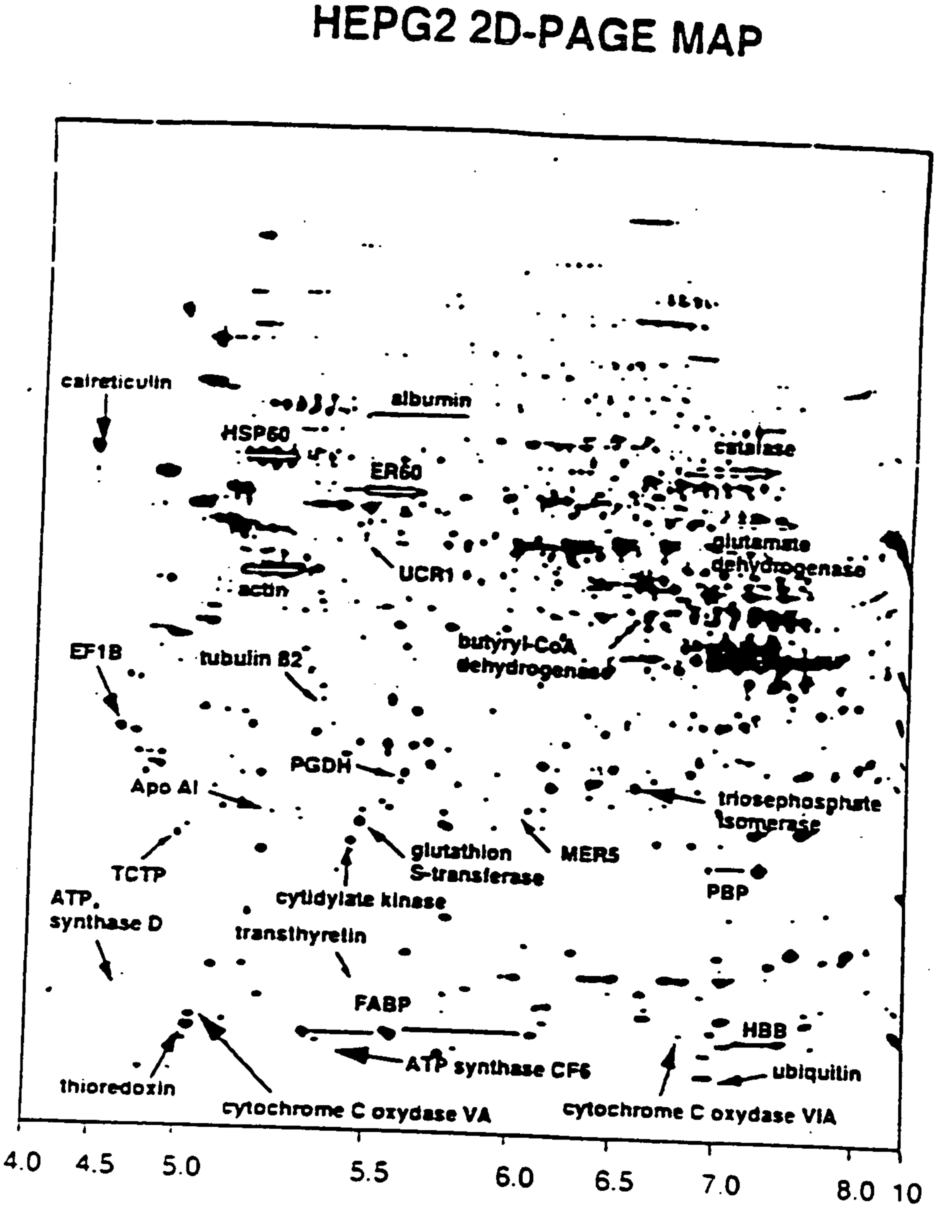


Figure 1. Two-dimensional gel electrophoresis map of a human hepatoblastoma-derived cell line. illustrating the very high resolution of the technique. The first dimensional separation (right to left of figure) was achieved using immobilised pH gradient electrophoresis of 4.0 to 10.0 units. The second dimension (top to bottom of figure) was SDS-PAGE using a 11%–14% acrylamide gradient, allowing separation in the molecular weight range 10–250 kDa. Proteins were visualised by silver staining. Arrows show proteins of known identity.

1992; Celis *et al*., 1993; Garrels and Franza, 1989; VanBogelen *et al*., 1992). Current protocols can resolve two to three thousand proteins from a complex sample on a single gel (*Figure 1*).

2-D GEL RESOLUTION AND REPRODUCIBILITY

A primary challenge of separating complex mixtures of proteins by 2-D gel electrophoresis has been to achieve high resolution and reproducibility. High resolution ensures that a maximum of protein species are separated, and high reproducibility is vital to allow comparison of gels from day to day and between research sites. These factors can be difficult to achieve.

Carrier ampholytes are a common means of isoelectric focusing for the first dimension of 2-D electrophoresis. Gels are usually focused to equilibrium to separate proteins in the pI range 4 to 8, and run in a non-equilibrium mode (NEPHGE) to separate proteins of higher pI (7 to 11.5) (O'Farrell, 1975; O'Farrell, Goodman and O'Farrell, 1977). Unfortunately, the use of carrier ampholytes in the isoelectric focusing procedure is susceptible to 'cathode drift', whereby pH gradients established by prefocusing of ampholytes slowly change with time (Righetti and Drysdale, 1973). Carrier ampholyte pH gradients are also distorted by high salt concentration of samples (Bjellqvist *et al.*, 1982), and by high protein load (O'Farrell, 1975). A further limitation is that iso electric focusing gels, which are cast and subject to electrophoresis in narrow glass tubes, need to be extruded by mechanical means before application to the second dimension – a procedure that potentially distorts the gel. Nevertheless, many of the above shortcomings can be avoided by loading small amounts of ^{14}C or ^{35}S radiolabelled samples (Garrels, 1989; Neidhardt *et al.*, 1989; Vandekerkhove *et al.*, 1990). High sensitivity detection is then achieved through use of fluorography or phosphorimaging plates (Bonner and Laskey, 1974; Johnston, Pickett and Barker, 1990; Patterson and Latter, 1993). However, this approach is only practicable for organisms or tissues that can be radiolabelled.

An alternative technique, which is becoming the method of choice for the first dimension separation of proteins, involves isoelectric focusing in immobilized pH gradient (IPG) gels (Bjellqvist *et al.*, 1982; Görg, Postel and Gunther, 1988; Righetti, 1990). Immobilized pH gradients are formed by the covalent coupling of the pH gradient into an acrylamide matrix, creating a gradient that is completely stable with time. IPG gels are usually poured onto a stiff backing film, which is mechanically strong and provides easy gel handling (Ostergren, Eriksson and Bjellqvist, 1988). The major advantages of IPG separations are that they do not suffer from cathodic drift, they allow focusing of basic and very acidic proteins to equilibrium, pH gradients can be precisely tailored (linear, stepwise, sigmoidal), and that separations over a very narrow pH range are possible (0.05 pH units per cm) (Righetti, 1990; Bjellqvist *et al.*, 1982, 1993a; Sinha *et al.*, 1990; Görg *et al.*, 1988; Gelfi *et al.*, 1987; Gunther *et al.*, 1988). However, it is not currently possible to use IPG gels to separate very basic proteins of isoelectric point greater than 10, although this is under development. Narrow pH range separations are useful to address problems of protein co-migration in complex samples, allowing 'zooming in' on regions of a gel (*Figure 2*). IPG gel strips are now commercially available, which begin to address the problems of intra- and inter-lab isoelectric focusing reproducibility.

There are two means of electrophoresis for the second dimension separation of proteins: vertical slab gels and horizontal ultrathin gels (Görg, Postel, and Gunther, 1988). Both are usually SDS-containing gradient gels of approximately 11% to 15% acrylamide, which separate proteins in the molecular mass range of 10 – 150kD. A stacking gel is not usually used with slab gels, but is necessary when using horizontal gel setups (Görg, Postel and Gunther, 1988). Comparisons have shown that there is little or no difference in the reproducibility of electrophoresis using either approach (Corbett *et al.*, 1994a), but commercially available vertical or horizontal precast gels will provide greater reproducibility for occasional users. For slab gel electrophoresis,

Fi
2 p
pr
Th
wi
pl
se
Pr

th
ca
dc
1c

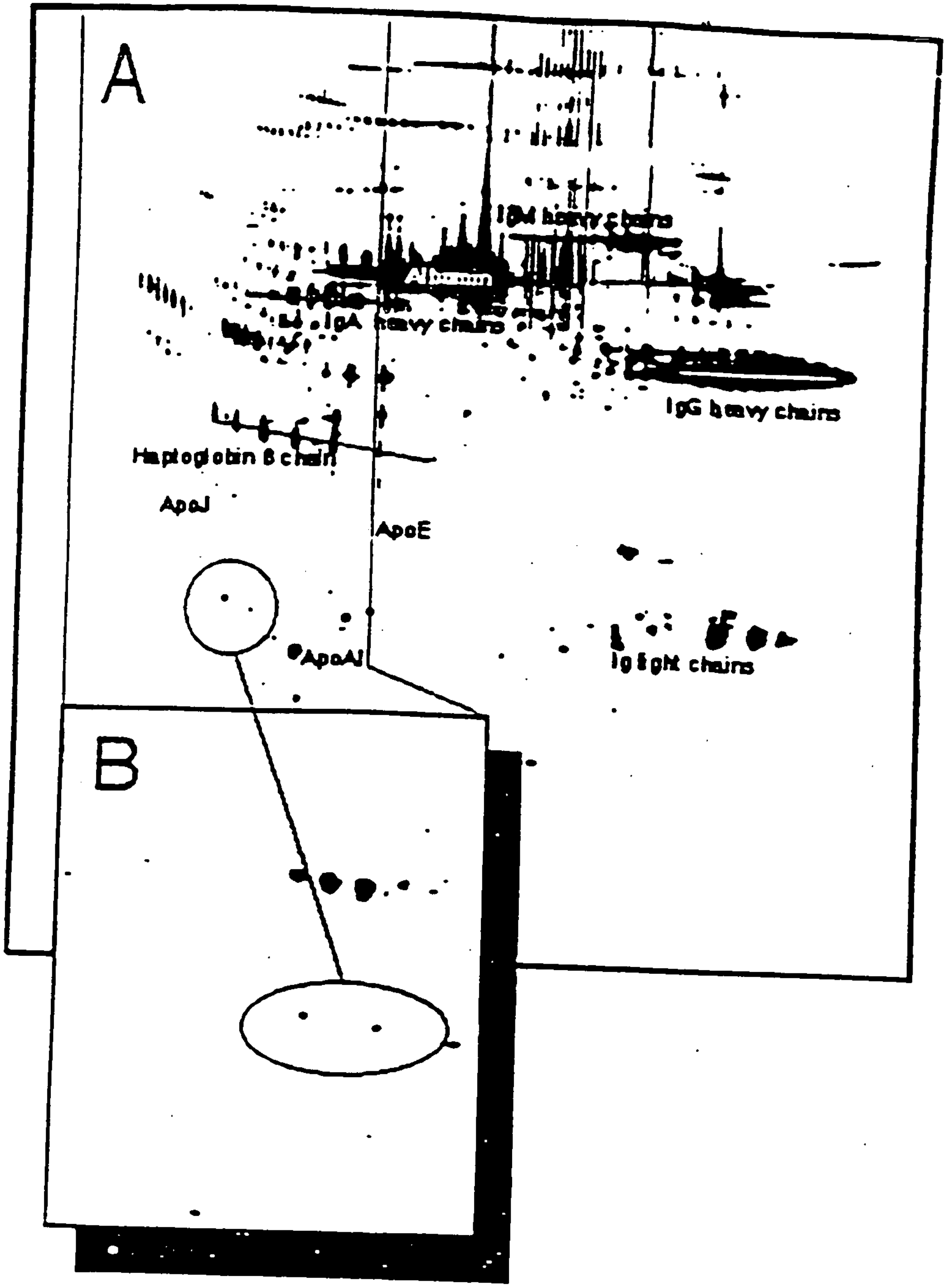


Figure 2. Two-dimensional gel electrophoresis allows 'zooming in' on areas of interest. Rings highlight 2 proteins common to each gel. (A) Wide pI range two dimensional electrophoresis map of human plasma proteins. First dimension separation was acheived using an immobilised pH gradient of 3.5 to 10.0 units. The second dimension was SDS-PAGE. Actual gel size was 16cm x 20cm, and proteins were visualised with silver staining. (B) Narrow pI range electrophoresis was used to 'zoom in' on a small region of the plasma map. The first dimension used a narrow range immobilised pH gradient of 4.2 to 5.2 units, and second dimension was SDS-PAGE. Micropreparative loading was used, and the gel blotted to PVDF. Proteins were visualised with amido black. Actual blot size was 16cm x 20cm

the use of piperazine diacrylyl as a gel crosslinker and the addition of thiosulfate in the catalyst system has been shown to give better resolution and higher sensitivity detection (Hochstrasser and Merril, 1988; Hochstrasser, Patchornik and Merril, 1988).

Notwithstanding the advances described above, there is an increasing demand to improve the reproducibility of 2-D electrophoresis to facilitate database construction and proteome studies. Harrington *et al.* (1993) explain that if a gel resolves 4000 protein spots, and there is 99.5% spot matching from gel to gel, this will produce 20 spot errors per gel. This amount of error, which might accumulate with each gel to gel comparison used in database construction, could produce an unacceptable degree of uncertainty in gel databases. To address these issues, partial automation of large 2-D gel separations has been undertaken (Nokihara, Morita and Kuriki, 1992; Harrington *et al.*, 1993). Although results are preliminary, spot to spot positional reproducibility in one study was found to be threefold improved over manual methods (Harrington *et al.*, 1993). It should be noted that small 2-D gel formats (50 × 43 mm) have been almost completely automated (Brewer *et al.*, 1986), although these are not generally used for database studies.

MICROPREPARATIVE 2-D GEL ELECTROPHORESIS

With the advent of affordable protein microcharacterisation techniques, including N-terminal microsequencing, amino acid analysis, peptide mass fingerprinting, phosphate analysis and monosaccharide compositional analysis, a new challenge for 2-D electrophoresis has been to maintain high resolution and reproducibility but to provide protein in sufficient quantities for chemical analysis (high nanogram to low microgram quantities of proteins per spot). This becomes difficult to achieve with very complex samples such as whole bacterial cells, as the initial protein load is divided among 2000 to 4000 protein species. Two approaches are used for producing amounts of material that can be chemically characterised. The first method is to run multiple gels, collect and pool the spots of interest, and subject them to concentration (Ji *et al.*, 1994; Walsh *et al.*, 1995; Rasmussen *et al.*, 1992). In this approach, the concentration process must also act as a purification step to remove accumulated electrophoretic contaminants such as glycine. A more elegant approach has been to exploit the high loading capacity of IPG isoelectric focusing. The high loading capacity of immobilised pH gradients was described early (Ek, Bjellqvist and Righetti, 1983), but has only recently been applied to 2-D electrophoresis (Hanash *et al.*, 1991; Bjellqvist *et al.*, 1993b). Up to 15 mg of protein can been applied to a single gel, yielding microgram quantities of hundreds of protein species. A further benefit of this approach is that proteins present in low abundance, which may not be visualised by lower protein loads, are more likely to be detected. The use of electrophoretic or chromatographic prefractionation techniques (Hochstrasser *et al.*, 1991a; Harrington *et al.*, 1992), followed by high loading of narrow-range IPG separations (Bjellqvist *et al.*, 1993b) provides a likely solution to studies on proteins present in low abundance.

Methods of protein detection

There are many means for detecting proteins from 2-D gels. The method used will be dictated by factors including protein load on gel (analytical or preparative), the purpose of the gel (for protein quantitation or for blotting and chemical characterisation), and the sensitivity required. The most common means of protein detection and their applications are shown in *Table 1*. Most detection methods have drawbacks, for

Table 1: Common stains for 2-D gels or blots and their applications.

Detection Method	Main applications	Unsuitable applications	Sensitivity	References
$[^{35}S]$ Met or ^{14}C radiolabelling and fluorography or phosphorimaging	Cell lines, cultured organisms	Samples that cannot be labelled	20 ppm of radiolabel in a spot	Garrels and Franza, 1989 Latham, Garrels and Solter, 1993
$[^{35}S]$thiourea silver	Extremely high sensitivity gel staining	Preparative 2-D, PVDF or NC membranes	0.4 ng protein on spot or band of gel	Wallace and Saluz, 1992a,b
Silver	Very high sensitivity gel staining, can be mono or polychromatic	Preparative 2-D, PVDF or NC membranes	4 ng protein on spot or band of gel	Rabilloud, 1992, Hochstrasser and Merril, 1988
Coomassie blue R-250	Staining of gels, staining of PVDF membranes before protein sequencing	Staining prior to direct mass determination from PVDF; amino acid analysis on PVDF; detection of some glycoproteins	40 ng protein on band or spot of gel	Strupat *et al.*, 1994; Gharahdaghi *et al.*, 1992; Goldberg *et al.*, 1988; Sanchez *et al.*, 1992
Colloidal gold	Staining NC membranes, staining PVDF before direct MALDI-TOF	Gels	60 × higher than coomassie	Yamaguchi and Asakawa, 1988; Eckerskorn *et al.*, 1992; Strupat *et al.*, 1994
Zinc imidazole	Reverse staining of gels or membranes; may be beneficial in MALDI-TOF of peptides	Where positive image is required	Higher than coomassie	Ortiz *et al.*, 1992; James *et al.*, 1993
Ponceau S and amido black	Staining higher protein loads on PVDF, for protein sequencing or amino acid analysis	Staining prior to direct mass determination from PVDF	100 ng protein on band or spot of gel	Sanchez *et al.*, 1992; Strupat *et al.*, 1994; Wilkins *et al.*, 1995
India ink	Staining of membrane-bound proteins; staining PVDF before direct MALDI-TOF	Gel staining, not quantitative from protein to protein	1–10 ng	Li *et al.*, 1989, Hughes, Mack and Hamparian, 1988, Strupat *et al.*, 1994
Stains-all	Staining to detect glycoproteins or Ca^{2+} binding proteins	General gel staining	100 ng protein on band or spot of gel	Campbell, MacLennan and Jorgensen, 1983; Goldberg *et al.*, 1988

PVDF = polyvinylidene difluoride, NC = nitrocellulose, MALDI-TOF = matrix assisted laser desorption ionisation time of flight mass spectrometry

example, some glycoproteins are not stained by coomassie blue (Goldberg *et al.*, 1988), and many organic dyes are unsuitable for protein detection on PVDF if samples are to be used for direct matrix-assited laser desorption ionisation mass spectrometry (Strupat *et al.*, 1994).

Although most means of protein detection give some indication of the quantities of protein present, in general they cannot be used for global quantitation. This is because no protein stain is able consistently to detect proteins over a wide range of concentrations, isoelectric points and amino acid compositions, and with a variety of post-translational modifications (Goldberg *et al.*, 1988; Li *et al.*, 1989). Furthermore, there are large differences in staining pattern when identical gels or blots are subjected to different stains, including amido black, imidazole zinc, india ink, ponceau S, colloidal gold, or coomassie blue (Tovey, Ford and Baldo, 1987; Ortiz *et al.*, 1992). The most common means of quantitating large numbers of proteins in a 2-D gel involves the radiolabelling of protein samples prior to electrophoresis, and protein quantitation based on fluorography and image analysis or liquid scintillation counting (Garrels, 1989; Celis and Olsen, 1994). However, proteins which do not contain methionine cannot be detected if only [^{35}S] methionine is used for labelling. Amino acid analysis of protein spots visualised by other techniques presents a likely means of protein quantitation for the future.

BLOTTING OF PROTEINS TO MEMBRANES

Electrophoretic blotting of proteins from two-dimensional polyacrylamide gels to membranes presents many options for protein identification and microcharacterisation which are not possible when proteins remain in gels. For example, when proteins are blotted to polyvinylidene difluoride (PVDF) membranes, they can be identified by N-terminal sequencing, amino acid analysis, or immunoblotting, or they may be subjected to endoproteinase digestion, monosaccharide analysis, phosphate analysis, or direct matrix-assisted laser desorption ionisation mass spectrometry (Matsudaira, 1987; Wilkins *et al.*, 1995; Jungblut *et al.*, 1994; Sutton *et al.*, 1995; Rasmussen *et al.*, 1994; Weizthandler *et al.*, 1993; Murthy and Iqbal, 1991; Eckerskorn *et al.*, 1992). It is possible to combine of some of these procedures on a single protein spot on a PVDF membrane (Packer *et al.*, 1995; Wilkins *et al.*, submitted; Weizthandler *et al.*, 1993). This is useful when minimal amounts of protein are available for analysis. These techniques will be explored in detail later in this review. Notwithstanding the above, there are some disadvantages associated with blotting of proteins to membranes. There is always loss of sample during blotting procedures (Eckerskorn and Lottspeich, 1993), and common protein detection methods are less sensitive or not applicable to membranes (*Table 1*), presenting difficulties for the analysis of low abundance proteins. Detailed discussion of the merits of available membranes and common blotting techniques can be found elsewhere (Eckerskorn and Lottspeich, 1993; Strupat *et al.*, 1994; Patterson, 1994).

2-D gel analysis, documentation, and proteome databases

Following protein electrophoresis and detection, detailed analysis of gel images is undertaken with computer systems. For proteome projects, the aim of this analysis is to catalogue all spots from the 2-D gel in a qualitative and if possible quantitative manner, so as to define the number of proteins present and their levels of expression. Reference gel images, constructed from one or more gels, form the basis of two-dimensional gel databases. These databases also contain protein spot identities and details of their post-translational modifications. 2-D gel databases are beginning to be linked to or integrated with comprehensive protein and nucleic acid databases (Neidhardt *et al.*, 1989; Simpson *et al.*, 1992; Appel *et al.*, 1994), and 'organism' databases, containing DNA sequence data, chromosomal map locations, reference 2-D gels and protein functional information for an organism, are becoming established as genome and proteome projects progress (VanBogelen *et al.*, 1992; Yeast Protein Database cited in Garrels *et al.*, 1994).

GEL IMAGE ANALYSIS AND REFERENCE GELS

After 2-D electrophoresis and protein visualisation by staining, fluorography or phosphorimaging, images of gels are digitised for computer analysis by an image scanner, laser densitomer, or charge-coupled device (CCD) camera (Garrels, 1989; Celis *et al.*, 1990a; Urwin and Jackson, 1993). All systems digitise gels with a resolution of 100 – 200 mm, and can detect a wide range of densities or shading (256 or more 'grey scales'). Following this, gel images are subjected to a series of manipulations to remove vertical and horizontal streaking and background haze, to detect spot positions and boundaries, and to calculate spot intensity (*Figure 3*). A standard spot (SSP) number, containing vertical and horizontal positional information, is assigned to each detected spot and becomes the protein's reference number. *Table 2* lists some notable software packages which process 2-D gel images.

Table 2: Some Software Packages for the Analysis of Gel Images.

Gel Image Analysis System	References*
ELSIE 4 & 5	Olsen and Miller, 1988; Wirth *et al.*, 1991; Wirth *et al.*, 1993
GELLAB I & II	Wu, Lemkin and Upton, 1993; Lemkin, Wu and Upton, 1993, Myrick *et al.*, 1993.
MELANIE I & II	Appel, *et al.* 1991, Hochstrasser *et al.* 1991b
QUEST I & II and PDQUEST	Garrels, 1989, Monardo *et al.*, 1994, Holt *et al.*, 1992, Celis *et al.*, 1990a,b
TYCHO & KEPLAR	Anderson *et al.*, 1984, Richardson, Horn and Anderson, 1994

* These references are not exhaustive, they include some references of use as well as authors of the system

As there are difficulties in the electrophoresis of samples with 100% reproducibility, reference gel images are often constructed from many gels of the same sample (Garrels and Franza, 1989; Neidhardt *et al.*, 1989). Since this involves the matching of 2000 to 4000 proteins from one gel to another, it presents a considerable challenge to image analysis systems. Matching of gels is usually initiated by an operator, who manually designates approximately 50 or so prominent spots as 'landmarks' on gels to be cross-matched. Proteins which match are then established around landmarks, using computer-based vector algorithms to extend the matching over the entire gel. Close to 100% of spots from complex samples can be matched by these methods, although different degrees of operator intervention may be required (Olsen and Miller, 1988; Lemkin and Lester, 1989; Garrels, 1989; Myrick *et al.*, 1993).

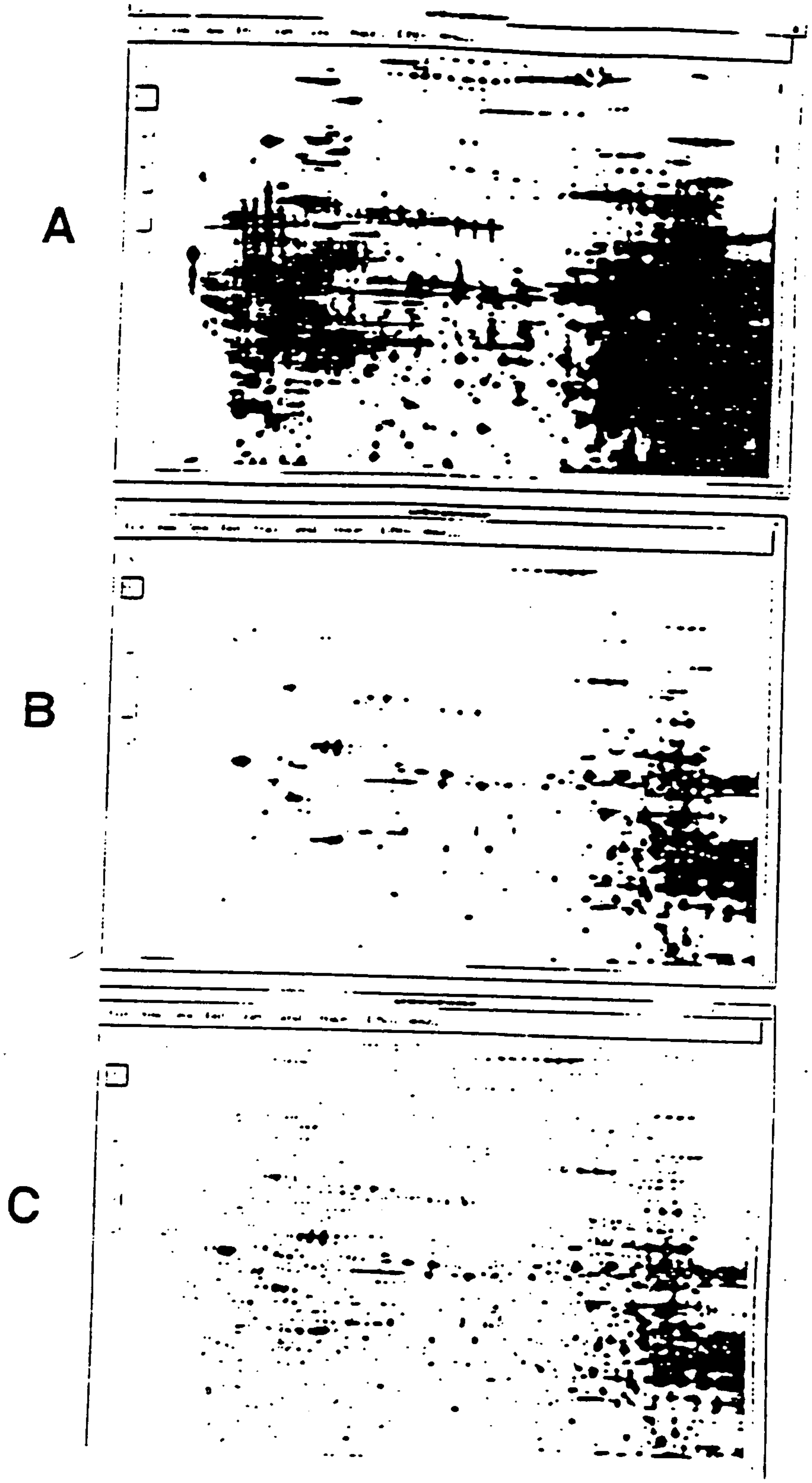


Figure 3. Computer processing of gel images. Shown is a wide pI range 2-D separation of human liver proteins, processed by Melanie software (Appel *et al.*, 1991). (A) Original gel image as captured by laser densitometer. (B) Gel image after processing to remove streaking and background. (C) Outline definition of all spots on the gel.

Exp
2-D
durin
are re
obtai
curve
MW
Boge
al., 1
to PV
et al.
prote
amino
carry
positi

SPOT

A ma
separ
to de
positi
of ma
perfo
minut
scintil
by re
protei
et al.,
Garre
be no
Limit
not ac
only e
an alte
Myric
1992.
Wh
and m
protei
their a
transf
regul
synth
(Lath.

CALCULATION OF PROTEIN ISOELECTRIC POINT AND MOLECULAR WEIGHT

Estimation of the isoelectric point (pI) and molecular weight (MW) of proteins from 2-D gels provides fundamental parameters for each protein, which are also of use during identification procedures (see following section). The pI and MW of proteins are recorded in 2-D gel databases. Accurate estimations of protein pI and MW can be obtained by using 20 or more known proteins on a reference map to construct standard curves of pI and molecular weight, which are then used to calculate estimated pI and MW of unknown proteins (Neidhardt *et al.*, 1989; Garrels and Franza, 1989; VanBogelen, Hutton and Neidhardt, 1990; Anderson and Anderson, 1991; Anderson *et al.*, 1991; Latham *et al.*, 1992). Alternatively, the MW of individual proteins blotted to PVDF can be determined very accurately by direct mass spectrometry (Eckerskorn *et al.*, 1992). Where immobilised pH gradients are used, the focusing position of proteins allows their pI to be measured within 0.15 units of that calculated from the amino acid sequence (Bjellqvist *et al.*, 1993c). It must be noted, however, that proteins carrying post-translational modifications may migrate to unexpected pI or MW positions during electrophoresis (Packer *et al.*, 1995).

SPOT QUANTITATION AND EXPRESSION ANALYSIS

A major challenge faced in proteome projects is the quantitative analysis of proteins separated by 2-D electrophoresis. The most accurate means of protein quantitation is to determine chemically the amount of each protein present by amino acid compositional analysis. However, the current method of choice for quantitative analysis of many proteins is to radiolabel samples with $[^{35}S]$ methionine or ^{14}C amino acids, perform the 2-D electrophoresis, and measure protein levels in disintegrations per minute (dpm) or units of optical density. Quantitation is achieved either by liquid scintillation counting, or by gel image analysis where spot densities are quantitated by reference to gel calibration strips containing known amounts of radiolabelled protein or against the integrated optical density of all spots visualised (Vandekerkhove *et al.*, 1990; Celis *et al.*, 1990b; Celis and Olsen, 1994; Garrels, 1989; Latham, Garrels and Solter, 1993; Fey *et al.*, 1994). All approaches effectively allow spots to be normalised against the total disintegrations per minute loaded onto the gel. Limitations that remain with radiolabelling methods are that absolute quantitation is not achieved because all proteins have varying amounts of any amino acid, and that only easily labelled samples can be investigated. Quantitative silver staining presents an alternative (Giometti *et al.*, 1991; Harrington *et al.*, 1992; Rodriguez *et al.*, 1993; Myrick *et al.*, 1993), which when undertaken with $[^{35}S]$thiourea (Wallace and Saluz, 1992 a,b) is of extremely high sensitivity.

When protein spots from samples prepared under different conditions are quantitated and matched from gel to gel, it becomes possible to examine changes and patterns in protein expression. Large scale investigation of up- and down-regulation of proteins, their appearance and disappearance, can be undertaken. For example, simian virus 40 transformed human keratinocytes were shown to have 177 up-regulated and 58 down-regulated proteins compared to normal keratinocytes (Celis and Olsen, 1994); detailed synthesis profiles of 1200 proteins have been established in 1 to 4 cell mouse embryos (Latham *et al.*, 1991, 1992); and 4 proteins out of 1971 were found to be markers for cadmium toxicity in urinary proteins (Myrick *et al.*, 1993). Complex global changes in protein expression as a result of gene disruptions have also been investigated (S. Fey and P. Mose-Larsen, Personal communication). Impressively, large gel sets showing protein expression under different conditions can be globally investigated using statistical methods that find groups of related objects within a set. For example, the REF52 rat cell line database, consisting of 79 gels from 12 experimental groups where each gel contains quantitative data for 1600 cross-matched proteins, has been analysed by cluster analysis (Garrels *et al.*, 1990). This revealed clusters of proteins that, for example, were induced or repressed similarly under simian virus 40 or adenovirus transformation, suggesting a common mechanism. Protein groups that were induced or repressed during culture growth to confluence were also found. It is obvious that the potential for investigation of cellular control mechanisms by these approaches is immense. It is equally clear that investigations of gene expression of this scale are currently technically impossible using nucleic-acid based techniques.

Table 3: Some proteome databases and their special features

Proteome database	Special features	References
E. coli gene-protein database	Gel spots linked with GenBank and Kohara clones; quantitative spot measurements under different growth conditions	VanBogelen and Neidhardt, 1991; VanBogelen *et al.*, 1992
Human heart databases	Identification of disease markers; two separate databases have been established	Baker *et al.*, 1992 Corbett *et al.*, 1994b Jungblut *et al.*, 1994
Human keratinocyte database	Extensive identifications; quantitative spot measurements of transformed cells; identification of disease markers	Celis *et al.*, 1990a Celis *et al.*, 1993 Celis and Olsen 1994
Mouse embryo database	Quantitative spot measurements through 1 to 4 cell stage	Latham *et al.*, 1991 Latham *et al.*, 1992
Mouse liver database (Argonne Protein Mapping Group)	Documents changes due to exposure to ionizing radiation and toxic chemicals	Giometti, Taylor and Tollaksen, 1992
Rat liver epithelial database	Detailed subcellular fractionation studies	Wirth *et al.*, 1991 Wirth *et al.*, 1993
Rat liver database	Extensive studies on regulation of proteins by drugs and toxic agents	Anderson and Anderson, 1991, Anderson *et al.*, 1992; Richardson, Horn and Anderson, 1994
REF 52 rat cell line database	Accessible via World Wide Web, quantitative spot measurements under different conditions	Garrels and Franza 1989 Boutell *et al.*, 1994
SWISS-2DPAGE containing human reference maps	Accessible via World Wide Web, completely integrated with SWISS-PROT and SWISS-3DIMAGE	Appel *et al.*, 1993 Hochstrasser *et al.*, 1992 Hughes *et al.*, 1993 Golaz *et al.*, 1993
Yeast Protein Database (YPD) and Yeast Electrophoretic Protein Database (YEPD)	Completely crossreferenced organism database; YPD has extensive information on over 3500 proteins; YEPD has many identifications	Garrels *et al.*, 1994

FEATURES OF PROTEOME DATABASES

Proteome projects rely heavily on computer databases to store information about all proteins expressed by an organism. 'Proteome databases' should contain detailed information of proteins already characterised elsewhere, as well as protein data from 2-D gels such as apparent pI and MW, expression level under different conditions, subcellular localisation, and information on post-translational modifications. Images of reference 2-D gels, showing protein SSP numbers and protein identifications, should also be included. Ideally, proteome databases should be accessible with Macintosh or IBM personal computers and easy to use. Some proteome databases and the areas they cover are listed in *Table 3*. Databases range from collections of annotated gels to large databases of images integrated with protein and nucleic acid sequence banks.

One example of an integrated proteome database is the suite of SWISS-PROT, SWISS-2DPAGE and SWISS-3DIMAGE databases (Appel *et al*., 1993; Appel *et al*., 1994; Appel, Bairoch and Hochstrasser, 1994; Bairoch and Boeckmann, 1994). The features of these three databases are listed in *Table 4*. SWISS-PROT, SWISS-2DPAGE and SWISS-3DIMAGE are accessible through the World Wide Web

Table 4: The SWISS-PROT, SWISS-2DPAGE and SWISS-3DIMAGE suite of crosslinked databases. All three databases are accessible through the World Wide Web, at URL address: http://expasy.hcuge.ch/

	SWISS-PROT	SWISS-2DPAGE	SWISS-3DIMAGE
Information	Text entries of sequence data: Citation information; taxonomic data. 38,303 entries in Release 29	2-D gel images of: human liver, plasma, HepG2, HepG2 secreted proteins, red blood cell, lymphoma, cerebrospinal fluid, macrophage like cell line, erythroleukemia cell, platelet	Collection of 330 3-D images of proteins
Annotations	Protein function, Post translational modifications, Domains; Secondary structure, Quaternary structure, Diseases associated with protein, Sequence conflicts	Gel images where protein is found; How protein identified, Protein pI and MW, protein number; normal and pathological variants	All annotation is available in SWISS-PROT
Cross-Referenced Databases	SWISS-2DPAGE SWISS-3DIMAGE EMBL, PIR, PDB, OMIM, PROSITE, Medline; Flybase; GCRDb, MaizeDB, WormPep, DictyDB	SWISS-PROT and all other databases accessible through SWISS-PROT	SWISS-PROT and all other databases accessible through SWISS-PROT
Other Features	Navigation to other SWISS databases achieved by selecting entries with computer mouse	Gel images show position of identified proteins, or region of gel where protein should appear	Mono and stereo images available. Images can be transferred to local computer image viewing programs

(Berners-Lee *et al.*, 1992), allowing any computer connected to the internet to access the stored information and images. Navigation within and between the three databases is seamless, as all potential crosslinks are highlighted as hypertext on the display and can be selected with a computer mouse. From these databases, detailed information about a protein, including amino acid sequence and known post-translational modifications, can be obtained, the precise protein spot it corresponds to on a reference gel image can be viewed if known, and the 3-D structure of the molecule can be seen if available. References to nucleic acid and other databases are also given to provide access to information stored elsewhere.

'Organism' databases, containing detailed protein and nucleic acid information about a species, are becoming common as genome and proteome projects progress. These differ from nucleic acid or protein sequence databases like GenBank or SWISS-PROT because they are image based, and contain information about chromosomal map positions, transcription of genes, and protein expression patterns. The *Escherichia coli* gene-protein database (VanBogelen, Hutton and Neidhardt, 1990; VanBogelen and Neidhardt, 1991, VanBogelen *et al.*, 1992), known as the ECO2DBASE, is one example. It contains gene and protein names, 2-D gel spot information (including pI and MW estimates, and spot identification), genetic information (GenBank or EMBL codes, chromosomal location, location on Kohara clones (Kohara, Akiyama, and Isono, 1987), transcription direction of genes), and protein regulatory information (level of protein expression under different growth regimes, member of regulon or stimulon). All entries in the ECO2DBASE are also cross-referenced to the SWISS-PROT database (Bairoch and Boeckmann, 1994). It is anticipated that organism databases will soon become a standard means of storing all available information about a particular species. However there is currently no consistent manner in which organism databases are assembled, which may hamper comparisons in the future.

Identification and characterisation of proteins from 2-D gels

The number of proteins identified on a 2-D reference map determines its usefulness as a research and reference tool. As most reference maps have only a small proportion of proteins identified, a major aim of current proteome projects is to screen many proteins from 2-D maps, in order to define them as 'known' in current nucleic acid and protein databases, or as 'unknown'. Protein identification assists in confirmation of DNA open reading frames, and provides focus for DNA sequencing projects and protein characterisation efforts by pointing to proteins that are novel. Since there may be 3000—4000 proteins from a single 2-D map that require identification, the challenge in protein screening is to identify proteins quickly, with a minimum of cost and effort.

Traditionally, proteins from 2-D gels have been identified by techniques such as immunoblotting, N-terminal microsequencing, internal peptide sequencing, comigration of unknown proteins with known proteins, or by overexpression of homologous genes of interest in the organism under study (Matsudaira, 1987; Rosenfeld *et al.*, 1992; VanBogelen *et al.*, 1992; Celis *et al.*, 1993; Honore *et al.*, 1993; Garrels *et al.*, 1994). Whilst these techniques are powerful identification tools, they are too expensive or time and labour intensive to use in mass screening programs. A hierarchical approach to mass protein identification has been recently suggested as an

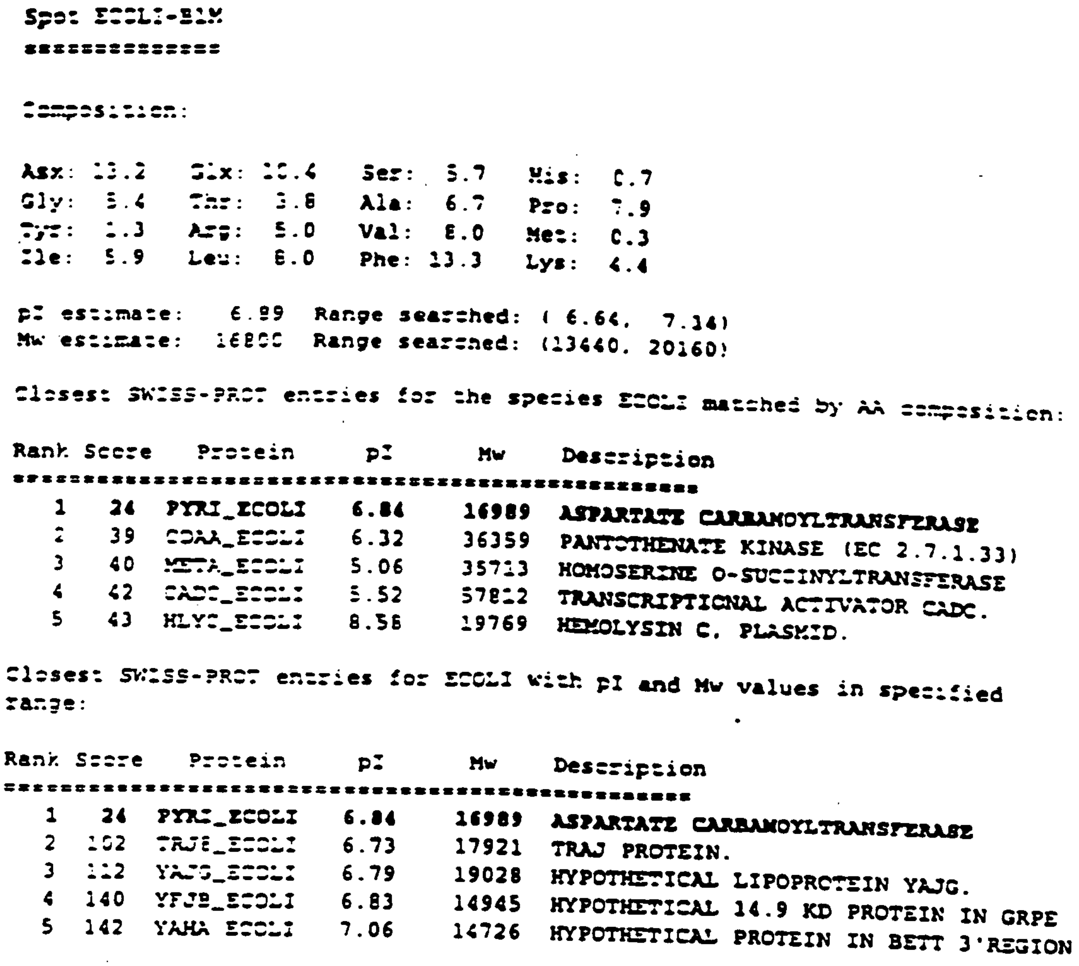
Table 5: Hierarchical analysis for mass screening of 2-D separated proteins blotted to membranes. Rapid and inexpensive techniques are used as a first step in protein identification, and slower, more expensive techniques are then used if necessary. Table modified from Wasinger *et al.* (1995).

Order	Identification technique	References
1	Amino acid analysis	Jungblut *et al.*, 1992, Shaw, 1993, Hobohm, Houthaeve and Sander, 1994, Jungblut *et al.*, 1994, Wilkins *et al.*, 1995
2	Amino acid analysis with N-terminal sequence tag	Wilkins *et al.*, submitted
3	Peptide-mass fingerprinting	Henzel *et al.*, 1993, Pappin, Hojrup and Bleasby, 1993, James *et al.*, 1993, Mann, Hojrup and Roepstorff, 1993, Yates *et al.*, 1993, Mortz *et al.*, 1994, Sutton *et al.*, 1995
4	Combination of amino acid analysis and peptide mass fingerprinting	Cordwell *et al.*, 1995, Wasinger *et al.*, 1995;
5	Mass spectrometry sequence tag	Mann and Wilm, 1994
6	Extensive N-terminal Edman microsequencing	Matsudaira, 1987
7	Internal peptide Edman microsequencing	Rosenfeld *et al.*, 1992; Hellman *et al.*, 1995,
8	Microsequencing by mass spectrometry (electrospray ionisation, post-source decay MALDI-TOF)	Johnson and Walsh, 1992
9	Ladder sequencing	Bartlet-Jones *et al.*, 1994

alternative to traditional approaches (*Table 5*; Wasinger *et al.*, 1995). This involves the use of rapid and cheap identification tools such as amino acid analysis and peptide mass fingerprinting as first steps in protein identification, followed by the use of slower, more expensive and time consuming identification procedures if necessary. In the construction of this hierarchy the analysis time, cost per sample and the complexity of the data created has been considered, as whilst some techniques require little machine time per sample, the analysis of data can be quite involved and time consuming. Amino acid analysis and peptide mass-fingerprinting based identification techniques in the hierarchy are discussed in detail below. For review of other protein identification techniques in *Table 5*, see Patterson (1994) and Mann (1995).

PROTEIN IDENTIFICATION BY AMINO ACID COMPOSITION

There has been a revival of interest in the use of amino acid composition for identification of proteins from 2-D gels after early work by Eckerskorn *et al.* (1988). This technique uses a protein's idiosyncratic amino acid composition profile in order to identify it by comparison with theoretical compositions of proteins in databases. The amino acid composition of proteins can be determined by differential metabolic radiolabelling and quantitative autoradiography after 2-D electrophoresis (Garrels *et al.*, 1994; Frey *et al.*, 1994), or by acid hydrolysis of membrane-blotted proteins and chromatographic analysis of the resulting amino acid mixture (Eckerskorn *et al.*, 1988; Tous *et al.*, 1989; Gharahdaghi *et al.*, 1992; Jungblut *et al.*, 1992; Wilkins *et al.*, 1995). As differential metabolic labelling experiments require X-ray film or phosphor-image plate exposures of up to 140 days, and can only be undertaken with easily radiolabelled samples, the technique is not as rapid or widely applicable as chromato-

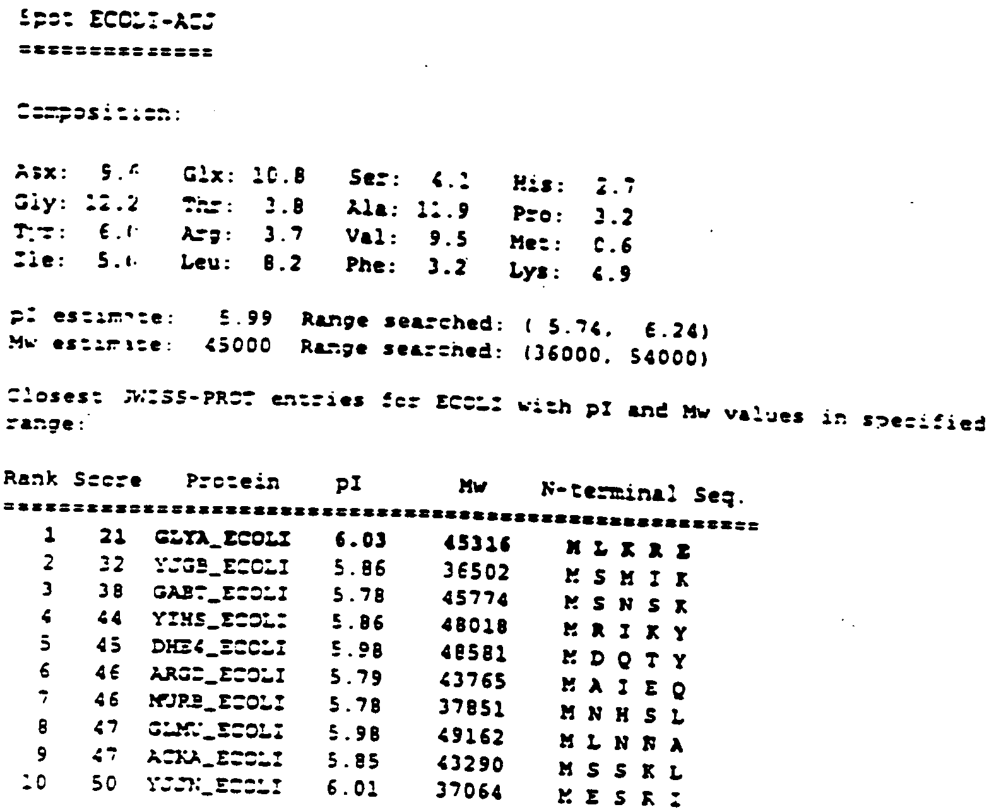


```
Spot ECOLI-B1M
==============

Composition:

Asx: 13.2   Glx: 10.4   Ser:  5.7   His:  0.7
Gly:  5.4   Thr:  3.8   Ala:  6.7   Pro:  7.9
Tyr:  1.3   Arg:  5.0   Val:  8.0   Met:  0.3
Ile:  5.9   Leu:  8.0   Phe: 13.3   Lys:  4.4

pI estimate:    6.99  Range searched: ( 6.64,  7.34)
Mw estimate:   16800  Range searched: (13440, 20160)

Closest SWISS-PROT entries for the species ECOLI matched by AA composition:

Rank Score   Protein      pI       Mw    Description
===================================================
   1   24   PYRI_ECOLI   6.84    16989  ASPARTATE CARBAMOYLTRANSFERASE
   2   39   COAA_ECOLI   6.32    36359  PANTOTHENATE KINASE (EC 2.7.1.33)
   3   40   META_ECOLI   5.06    35713  HOMOSERINE O-SUCCINYLTRANSFERASE
   4   42   CADC_ECOLI   5.52    57822  TRANSCRIPTIONAL ACTIVATOR CADC.
   5   43   HLYC_ECOLI   8.58    19769  HEMOLYSIN C, PLASMID.

Closest SWISS-PROT entries for ECOLI with pI and Mw values in specified
range:

Rank Score   Protein      pI       Mw    Description
===================================================
   1   24   PYRI_ECOLI   6.84    16989  ASPARTATE CARBAMOYLTRANSFERASE
   2  102   TRJE_ECOLI   6.73    17921  TRAJ PROTEIN.
   3  112   YAJG_ECOLI   6.79    19028  HYPOTHETICAL LIPOPROTEIN YAJG.
   4  140   YFJB_ECOLI   6.83    14945  HYPOTHETICAL 14.9 KD PROTEIN IN GRPE
   5  142   YAHA ECOLI   7.06    14726  HYPOTHETICAL PROTEIN IN BETT 3'REGION
```

Figure 4. Computer printout from ExPASy server where the empirical amino acid composition, estimated pI and MW of a protein from a 2-D reference map of *E. coli* were matched against all entries in SWISS-PROT for *E. coli*. The correct identification, aspartate carbamoyltransferase, is shown in bold. Low scores indicate a good match. Note how matching within a defined pI and MW range (lower set of proteins) has greatly increased the score difference between the first and second ranking proteins. This score difference gives high confidence in the identification, and is only observed where the top ranking protein is the correct identification (Wilkins *et al.*, 1995).

graphy-based analysis. Proteins blotted to PVDF membranes can be hydrolysed in 1 h at 155°C, amino acids extracted in a single brief step, and each sample automatically derivatised and separated by chromatography in under 40 minutes (Wilkins *et al.*, 1995; Ou *et al.*, 1995). In this manner, one operator can routinely analyse 100 proteins per week on one HPLC unit. This technology lends itself to automation, and it is anticipated that instruments with even greater sample throughput will be developed. When proteins have been prepared by micropreparative 2-D electrophoresis (Hanash *et al.*, 1991; Bjellqvist *et al.*, 1993b), blotted to a PVDF membrane and stained with amido black, any visible protein spot is of sufficient quantity for amino acid analysis (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995; Wilkins *et al.*, 1995).

After the amino acid composition of a protein has been determined, computer programs are used to match it against the calculated compositions of proteins in databases (Eckerskorn *et al.*, 1988; Sibbald, Sommerfeldt and Argos, 1991; Jungblut *et al.*, 1992; Shaw, 1993; Hobohm, Houthaeve and Sander, 1994; Wilkins *et al.*, 1995). Matching is usually done with only 15 or 16 amino acids, as cysteine and



```
Spot ECOLI-AC3
==============

Composition:

Asx:  9.6   Glx: 10.8   Ser:  4.1   His:  2.7
Gly: 12.2   Thr:  3.8   Ala: 11.9   Pro:  3.2
Tyr:  6.0   Arg:  3.7   Val:  9.5   Met:  0.6
Ile:  5.6   Leu:  8.2   Phe:  3.2   Lys:  4.9

pI estimate:    5.99  Range searched: ( 5.74,  6.24)
Mw estimate:   45000  Range searched: (36000, 54000)

Closest SWISS-PROT entries for ECOLI with pI and Mw values in specified
range:

Rank Score   Protein     pI       Mw    N-terminal Seq.
=======================================================
  1    21  GLYA_ECOLI   6.03    45316    M L K R E
  2    32  YJGB_ECOLI   5.86    36502    M S M I K
  3    38  GABT_ECOLI   5.78    45774    M S N S K
  4    44  YIHS_ECOLI   5.86    48018    M R I K Y
  5    45  DHE4_ECOLI   5.98    48581    M D Q T Y
  6    46  ARGD_ECOLI   5.79    43765    M A I E Q
  7    46  MURB_ECOLI   5.78    37851    M N H S L
  8    47  GLMU_ECOLI   5.98    49162    M L N N A
  9    47  ACKA_ECOLI   5.85    43290    M S S K L
 10    50  YJJN_ECOLI   6.01    37064    M E S R I
```

Figure 5. A PVDF protein spot from an *E. coli* 2-D reference map was sequenced for 4 cycles, and the same sample then subject to amino acid analysis. The N-terminal sequence was M L K R. When the amino acid composition of the spot, as well as estimated pI and MW, were matched against all entries in SWISS-PROT for *E. coli*, the above list of best matches was produced. N-terminal sequences are from SWISS-PROT for those entries. The top ranking identification of serine hydroxymethyltransferase (bold) did not show a large score difference between the first and second ranking proteins, giving little confidence in this being the correct protein identification. However, the sequence tag (M L K R) confirmed the identity of the protein as serine hydroxymethyltransferase.

tryptophan are destroyed during hydrolysis, asparagine and glutamine are deamidated to their corresponding acids, and proline is not quantitated in some analysis systems. The computer programs produce a list of best matching proteins, which are ranked by a score that indicates the match quality. Some programs allow matching to be restricted to specific 'windows' of MW and pI (Hobohm, Houthaeve and Sander, 1994; Wilkins *et al.*, 1995), and to protein database entries for one species (Jungblut *et al.*, 1992; Wilkins *et al.*, 1995). The use of such restrictions increases the power of matching. An example of protein identification by amino acid composition is shown in *Figure 4*. To date, amino acid composition has been used to identify proteins from reference maps of *Spiroplasma melliferum, Mycoplasma genitalium, E. coli, Saccharomyces cerevisiae, Dictyostelium discoideum*, human sera, human heart, human lymphocyte, and mouse brain (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995; Wilkins *et al.*, 1995; Jungblut *et al.*, 1992, 1994; Garrels *et al.*, 1994; Frey *et al.*, 1994).

PROTEIN IDENTIFICATION BY AMINO ACID COMPOSITION AND N-TERMINAL SEQUENCE TAG

When samples from 2-D gels are not unambiguously identified by amino acid composition, pI and MW, often the correct identification of that protein is amongst the top rankings of the list (Hobohm, Houthaeve and Sander, 1994; Cordwell *et al.*, 1995; Wilkins *et al.*, 1995). Taking advantage of this observation, we have used the mass spectrometry 'sequence tag' concept (Mann and Wilm, 1994) in developing a combined Edman degradation and amino acid analysis approach to protein identification (Wilkins *et al.*, submitted). This involves the N-terminal sequencing of PVDF-blotted proteins by Edman degradation for 3 or 4 cycles to create a 'sequence tag', following which the same sample is used for amino acid analysis. As only a few amino acids are removed from the protein, its composition is not significantly altered. Furthermore, since only a small amount of protein sequence is required, fast but low repetitive yield Edman degradation cycles can be used. Modifications to current procedures should allow 3 cycles to be completed in 1 h, thereby allowing the screening of 100 or more proteins per week on one automated, multi-cartridge sequenator. Amino acid composition, pI and MW of proteins are matched against databases as described above, and N-terminal sequences of best matching proteins are checked with the 'sequence tag' to confirm the protein identity (*Figure 5*). This technique will be less useful when proteins are N-terminally blocked, but as only a few N-terminal amino acids are susceptible to the acetyl, formyl, or pyroglutamyl modifications that cause blockage, this may itself provide useful information for sequence tag identification. A strength of N-terminal sequence tag and amino acid composition protein identification is that data generated are quickly and easily interpreted.

PROTEIN IDENTIFICATION BY PEPTIDE MASS FINGERPRINTING

Techniques for the identification of proteins by peptide mass fingerprinting have recently been described (Henzel *et al.*, 1993; Pappin, Hojrup and Bleasby, 1993; James *et al.*, 1993; Mann, Hojrup and Roepstorff, 1993; Yates *et al.*, 1993; Mortz *et al.*, 1994; Sutton *et al.*, 1995). This involves the generation of peptides from proteins using residue-specific enzymes, the determination of peptide masses, and the matching of these masses against theoretical peptide libraries generated from protein sequence databases. As proteins have different amino acid sequences, their peptides should produce characteristic 'fingerprints'.

The first step of peptide mass fingerprinting is protein digestion. Proteins within the gel matrix or bound to PVDF can be enzymatically digested *in situ*, although *in situ* gel digests are reported to produce more enzyme autodigestion products, which complicate subsequent peptide mass analysis (James *et al.*, 1993; Rasmussen *et al.*, 1994; Mortz *et al.*, 1994). The enzyme of choice for digestion is currently trypsin (of modified sequencing grade), but other enzymes (Lys-C or *S. aureus* V8 protease) have also been used (Pappin, Hojrup and Bleasby, 1993). To maximise the number of peptides obtained, it is desirable for protein samples to be reduced and alkylated prior to digestion (Mortz *et al.*, 1994; Henzel *et al.*, 1993). This ensures that all disulfide bonds of the protein are broken, and produces protein conformations that are more amenable to digestion. Surprisingly, chemical digestion methods such as cyanogen bromide (methionine specific), formic acid (aspartic acid specific), and 2-(2'-nitrophenylsulfenyl)-3-methyl-3'-bromoindolenine (tryptophan specific) have not been explored as means of peptide production for mass fingerprinting, even though they are rapid and may circumvent some problems associated with enzyme digestions

(Nikodem and Fresco, 1979; Crimmins *et al.*, 1990; Vanfleteren *et al.*, 1992).

After proteins are digested, peptide masses are determined by mass spectrometry. Direct analysis of peptide mixtures can be achieved by electrospray ionisation mass spectrometry, plasma desorption mass spectrometry, or matrix assisted laser desorption ionization (MALDI) mass spectrometry techniques. MALDI is preferable because of its higher sensitivity and greater tolerance to contaminating substances from 2-D gels (James *et al.*, 1993; Mortz *et al.*, 1994; Pappin, Hojrup and Bleasby, 1993). Furthermore, recent modifications to sample preparation methods have largely solved early difficulties experienced with the calibration of MALDI spectra (Mortz *et al.*, 1994; Vorm and Mann, 1994; Vorm, Roepstorff and Mann, 1994). The high sensitivity of mass spectrometry allows a small fraction of a digest of a 1µg protein spot to be used for analysis, and analysis itself is complete in a few minutes.

A major challenge associated with peptide mass fingerprinting is data interpretation prior to computer matching against libraries of theoretical peptide digests. Spectra must be examined carefully to determine which peaks represent peptide masses of interest, as there are often enzyme autodigestion products and contaminating substances present (Henzel *et al.*, 1993; Mortz *et al.*, 1994; Rasmussen *et al.*, 1994). Furthermore, if protein alkylation and reduction has not been undertaken prior to protein digestion, peptide sequence coverage may be poor (40% to 70%), with some masses present representing disulfide bonded peptides originally present in the protein (Mortz *et al.*, 1994). For eukaryotes, a serious issue is the alteration of peptide masses by the presence of post-translational modifications (*Table 6*). The mass of the unmodified peptide alone can be very difficult to determine. Two artifactual modifications introduced by electrophoresis, an acrylamide adduct to cysteine and the oxidation of methionine, are also known to alter peptide masses (le Maire *et al.*, 1993; Hess *et al.*, 1993).

Table 6: Masses of some common post-translational modifications. Peptides carrying post-translational modifications complicate data analysis for peptide mass fingerprinting protein identification. This is especially so for protein glycosylation, which involves many different combinations of the hexosamines, hexoses, deoxyhexoses, and sialic acid

Post-translational modification	Mass change
Acetylation	+ 42.04
*Acrylamide adduct to cysteine	+71.00
Carboxylation of Asp or Glu	+ 44.01
Deamidation of Asn or Gln	+ 0.98
Disulfide bond formation	- 2.02
Deoxyhexoses (Fuc)	146.14
Formylation	+ 28.01
Hexosamines (GlcN, GalN)	+ 161.16
Hexoses (Glc, Gal, Man)	+ 162.14
Hydroxylation	+ 16.00
N-acetylhexosamines (GlcNAc, GalNAc)	+ 203.19
*Oxidation of Met	+ 16.00
Phosphorylation	+ 79.98
Pyroglutamic acid formed from Gln	-17.03
Sialic acid (NeuNAc)	+ 291.26
Sulfation	+ 80.06

Table modified from Finnigan LASERMAT application data sheet 5.
Asterisk * shows modifications that can arise artifactually from the 2-D electrophoresis process

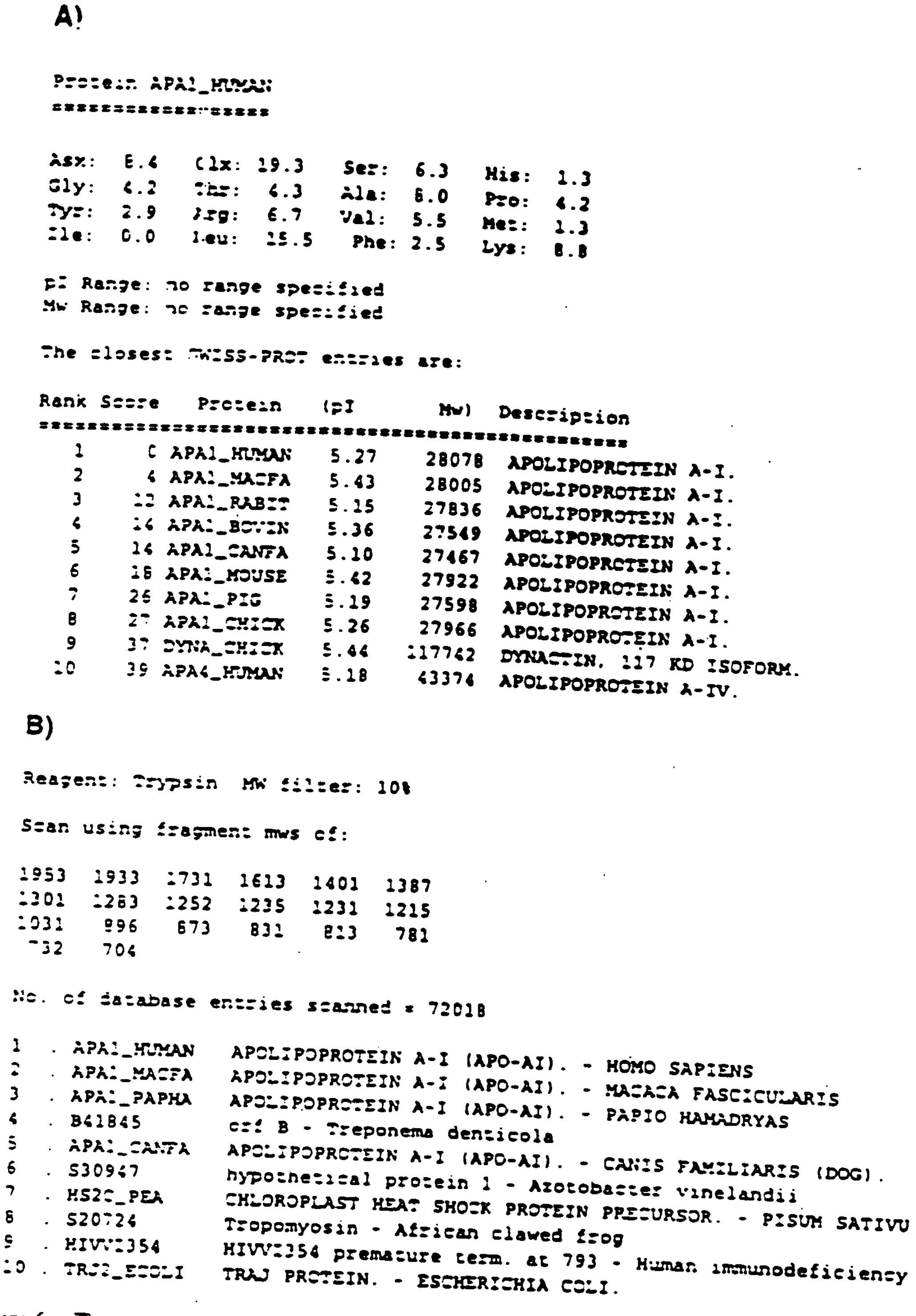
A number of computer programs are available for matching peptide masses against databases (reviewed in Cottrell, 1994). Matching is usually undertaken in an interactive manner, whereby peaks of mass 500–3000 Da are selected and matched under various search parameters including MW of protein, mass accuracy of peptides, and number of missed enzyme cleavages allowed (Henzel *et al.*, 1993; Mortz *et al.*, 1994; Rasmussen *et al.*, 1994). The correct protein identity is the protein which has the most peptide masses in common with the unknown sample. Identities have been established with as few as three peptides, but unambiguous identification is thought to require a mass spectrometric map covering most peptides of the protein (Mortz *et al.*, 1994; Yates *et al.*, 1993). To date, peptide mass fingerprinting of proteins has been undertaken from the human myocardial protein and keratinocyte maps, from an *E. coli* 2-D gel, and from reference maps of *Spiroplasma melliferum* and *Mycoplasma genitalium* (Sutton *et al.*, 1995; Rasmussen *et al.*, 1994; Henzel *et al.*, 1993; Cordwell *et al.*, 1995, Wasinger *et al.*, 1995), although the technique is most powerful when used in combination with another protein identification technique (Rasmussen *et al.*, 1994; Cordwell *et al.*, 1995).

MASS SPECTROMETRY SEQUENCE TAGGING

An extension of peptide mass fingerprinting has recently been described, called peptide sequence tagging (Mann and Wilm, 1994; Mann, 1995). This uses tandem mass spectrometry (MS/MS) to initially determine the mass of peptides, then subject them to fragmentation by collision with a gas, and finally determine the mass of fragments. The resulting spectra gives information about a peptide's amino acid sequence. The fragmentation masses of peptides can rarely be used to assign a complete sequence, but it usually allows a short 'sequence tag' of 2 or 3 amino acids to be determined. This sequence tag and the original peptide mass is matched by computer against a database, providing a likely identity of the peptide and the protein it came from. The major drawback for this technique as a mass screening tool is the complexity of the mass data generated and the high level of expertise required for its interpretation. Nevertheless, it represents a useful new protein identification method which greatly increases the power of peptide mass fingerprinting protein identification.

Cross-species protein identification

Protein sequence databases continue to grow at a rapid rate, yet it is not widely appreciated that close to 90% of all information contained in current protein databases comes from only 10 species (A. Bairoch, Pers. Comm.). Fortunately, this information can be used to study proteomes of organisms that are poorly defined at the molecular level, via 2-D electrophoresis and 'cross-species' protein identification (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995). This approach allows proteins from reference maps of many different species to be identified without the need for the corresponding genes to be cloned and sequenced. This is particularly true for 'housekeeping' proteins, such as enzymes involved in glycolysis, DNA manipulation and protein manufacture, which are highly conserved across species boundaries. Proteins that cannot be identified across species boundaries can then become the focus of further protein characterisation and DNA sequencing efforts.



A)

```
Protein APA1_HUMAN
==================

Asx:  8.4   Glx: 19.3   Ser:  6.3   His:  1.3
Gly:  4.2   Thr:  4.3   Ala:  8.0   Pro:  4.2
Tyr:  2.9   Arg:  6.7   Val:  5.5   Met:  1.3
Ile:  0.0   Leu: 15.5   Phe:  2.5   Lys:  8.8

pI Range: no range specified
Mw Range: no range specified

The closest SWISS-PROT entries are:

Rank Score   Protein    (pI       Mw)  Description
===================================================
  1      0 APA1_HUMAN   5.27     28078  APOLIPOPROTEIN A-I.
  2      4 APA1_MACFA   5.43     28005  APOLIPOPROTEIN A-I.
  3     12 APA1_RABIT   5.15     27836  APOLIPOPROTEIN A-I.
  4     14 APA1_BOVIN   5.36     27549  APOLIPOPROTEIN A-I.
  5     14 APA1_CANFA   5.10     27467  APOLIPOPROTEIN A-I.
  6     18 APA1_MOUSE   5.42     27922  APOLIPOPROTEIN A-I.
  7     26 APA1_PIG     5.19     27598  APOLIPOPROTEIN A-I.
  8     27 APA1_CHICK   5.26     27966  APOLIPOPROTEIN A-I.
  9     37 DYNA_CHICK   5.44    117742  DYNACTIN, 117 KD ISOFORM.
 10     39 APA4_HUMAN   5.18     43374  APOLIPOPROTEIN A-IV.
```

B)

```
Reagent: Trypsin  MW filter: 10%

Scan using fragment mws of:

1953  1933  1731  1613  1401  1387
1301  1283  1252  1235  1231  1215
1031   996   873   831   813   781
 732   704

No. of database entries scanned = 72018

1  . APA1_HUMAN   APOLIPOPROTEIN A-I (APO-AI). - HOMO SAPIENS
2  . APA1_MACFA   APOLIPOPROTEIN A-I (APO-AI). - MACACA FASCICULARIS
3  . APA1_PAPHA   APOLIPOPROTEIN A-I (APO-AI). - PAPIO HAMADRYAS
4  . B41845       orf B - Treponema denticola
5  . APA1_CANFA   APOLIPOPROTEIN A-I (APO-AI). - CANIS FAMILIARIS (DOG).
6  . S30947       hypothetical protein 1 - Azotobacter vinelandii
7  . HS2C_PEA     CHLOROPLAST HEAT SHOCK PROTEIN PRECURSOR. - PISUM SATIVU
8  . S20724       Tropomyosin - African clawed frog
9  . HIVVI354     HIVVI354 premature term. at 793 - Human immunodeficiency
10 . TRAJ_ECOLI   TRAJ PROTEIN. - ESCHERICHIA COLI.
```

Figure 6. Theoretical cross-species matching of human apolipoprotein A-I by amino acid composition and tryptic peptides. When an unknown protein is analysed, best ranking proteins from both techniques can be compared. If the same protein type is observed in both lists, there is high confidence in this being the identity of the unknown molecule (Cordwell *et al.*, 1995). (A) Output of ExPASy server (Appel, Bairoch and Hochstrasser, 1994) where the true amino acid composition of apolipoprotein A-I was matched against all entries in the SWISS-PROT database, without pI or MW windows. Seven of the top 10 matching proteins were apolipoprotein A-I of different species. (B) Output of MOWSE peptide mass fingerprinting program (Pappin, Hojrup and Bleasby, 1993) where true tryptic peptides of human apolipoprotein A-I were matched against the OWL database, using MW window of 10%. Four of the top ten matching proteins were apolipoprotein A-I from different species.

Rapid cross-species identification of proteins from 2-D reference maps can be undertaken with amino acid composition or peptide mass fingerprinting methods (*Figure 6*), but these techniques alone may not identify proteins unambiguously when phylogenetic cross-species distances are great or analysis data is of poor quality (Yates *et al.*, 1993; Shaw, 1993; Cordwell *et al.*, 1995). However, very high confidence in protein identities can be achieved when lists of best-matching proteins generated by both techniques are compared (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995). The correct identification is found when the same protein is ranked highly in lists of best matches generated by both techniques. This method has allowed approximately 120 proteins from the reference map of the mollicute *Spiroplasma melliferum*, representing approximately one quarter of the proteome, to be confidently identified by reference to protein information from other species (S. Cordwell, Personal Communication). When cross-species protein identification is to be undertaken, it should be noted that the molecular weight of a protein type across species is usually highly conserved, but that protein pI can vary by more than 2 units (Cordwell *et al.*, 1995). Accurate molecular weight determination by direct mass spectrometry of proteins blotted to PVDF (Eckerskorn *et al.*, 1992) should therefore be a useful additional parameter for cross-species protein identification.

CHARACTERISATION OF POST-TRANSLATIONAL MODIFICATIONS

Many proteins are modified after translation. Such post-translational modifications, including glycosylation, phosphorylation, and sulfation (see *Table 6*), are usually necessary for protein function or stability. Some abnormal modifications are associated with disease (Duthel and Revol, 1993; Ghosh *et al.*, 1993; Yamashita *et al.*, 1993). In proteome studies, post-translational modifications can be examined on all proteins present, or on individual spots. Studies on all proteins provide an indication of which proteins may carry a certain type of modification. For example, 2-D gel analysis of cell cultures grown in the presence of [^{3}H] mannose or [^{32}P] phosphate gives an indication of which proteins carry glycans containing mannose, and which proteins are phosphorylated (Garrels and Franza, 1989). Lectin binding studies of 2-D gels blotted to PVDF or nitrocellulose provide information on the saccharides, if any, that are carried by proteins present (Gravel *et al.*, 1994).

When individual proteins of interest carrying post-translational modifications have been found, micropreparative 2-D electrophoresis can be used to purify them in microgram quantities (Hanash *et al.*, 1991; Bjellqvist *et al.*, 1993b). If protein isoforms of similar MW and pI are to be studied, focusing with narrow range pI gradients (1 pH unit) can provide greater separation and resolution. After electrophoresis, the type and degree of protein phosphorylation can be investigated (Murthy and Iqbal, 1991; Gold *et al.*, 1994), monosaccharide composition can be determined (Weitzhandler *et al.*, 1993; Packer *et al.*, 1995), and the structure and exact site of glycoamino acids can be investigated by either Edman degradation based techniques or by mass spectrometry (Pisano *et al.*, 1993; Huberty *et al.*, 1993; Carr, Huddleston and Bean, 1993). With further development of rapid techniques, investigation of phosphorylation and monosaccharides by chromatographic or mass spectrometric means is likely to become a routine step in the characterisation of post-translational modifications of proteins from reference maps.

The status of proteome projects

Many technical aspects of proteome research have already been discussed in this review, but an overview of the status of proteome projects has not yet been presented. Advances in proteome projects will initially rely on progress in genome sequencing initiatives, to enable an identity, amino acid sequence, or function to be assigned to each protein spot. *Table 7* shows genome size, proteome size, and the number of proteins already defined for a number of model organisms. This indicates that whilst genome sequencing programs for *E. coli* and *S. cerevisiae* are advanced, the massive size of some other genomes (and especially the human genome) means that their complete nucleotide sequences are unlikely to be available for many years. Because of this, 2-D reference maps and proteome projects of single cell organisms like *Mycoplasma sp.*, *E. coli* and *S. cerevisiae* will be the most detailed (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995; Vanbogelen *et al.*, 1992; Garrels *et al.*, 1994), and complete maps of other organisms will take longer to construct. However, the use of cross-species protein identification techniques will allow proteomes of many prokaryotes and simple eukaryotes to be partially defined in reference to *E. coli* and *S. cerevisiae*.

Table 7: Estimated genome size, estimated proteome size, number of protein sequences in SWISS-PROT Release 31 (March, 1995), and approximate number of proteins of known identity on 2-D reference maps for some model organisms. Genome size data from Smith (1994), and total protein data from Bird (1995). Genome sequencing projects of *E. coli* and *S. cerevisiae* will probably be complete in 1996.

Species Name	Haploid genomesSize (million bp)	Estimated proteome size (total proteins)	Protein entries in SWISS PROT	Proteins annotated on 2-D Maps
Mycoplasma species	0.6–0.8	400–600	100	> 100
Escherichia coli	4.8	4000	3170	> 300
Saccharomyces cerevisiae	13.5	6000	3160	> 100
Dictyostelium discoideum	70	12500	204	-
Arabidopsis thaliana	70	14000	270	-
Caenorhabditis elegans	80	17800	703	-
Homo sapiens	2900	60000–80000	3326	> 1000

The study of vertebrate proteomes and vertebrate development is a phenomenal undertaking in comparison to the investigation of single cell organisms. This is because vast numbers of proteins are developmentally expressed, each body tissue has hundreds of unique proteins, and there are numerous tissue types. However, it is estimated that at least 35% of proteins in vertebrate cells will be conserved from tissue to tissue, constituting the 'housekeeping' proteins (Bird, 1995), with the remainder of proteins constituting a set that are specific to a cell type. Providing that standardised electrophoretic conditions are used, reference maps from many tissues of one organism can be superimposed in gel databases (e.g. Hochstrasser *et al.*, 1992). This accelerates the definition of the 'housekeeping' proteins, as well as sets of proteins that are unique to different tissue types. Such studies may, however, be complicated by post-translational modifications, which can differ on the same gene product in different tissues. Proteins that remain unknown after identification procedures will be useful in providing focus for nucleic acid sequencing initiatives.

FUTURE DIRECTIONS OF PROTEOME PROJECTS

This review has described recent advances in the area of proteome research. It has illustrated how new developments of older techniques (2-D electrophoresis and amino acid analysis) as well as the applications of new technology (mass spectrometry) have greatly widened the choice of tools the biologist and protein chemist has for the separation, identification and analysis of complex mixtures of proteins. This has made possible the establishment of detailed reference maps for organisms, which are becoming the method of choice for the definition of tissues or whole cells, and the investigation of gene expression therein.

Proteome projects are already impacting on the dogma of molecular biology that DNA sequence constitutes the definition of an organism. For example, the proteomes of different tissues of a single organism are often significantly different. Similarly, cross-species identification of proteins (for example the identification of proteins from *Candida albicans* by comparison with *S. cerevisiae*) can open up studies on organisms that are poorly molecularly defined. As cross-species identification can proceed at a pace orders of magnitude faster than a genome project in terms of defining the gene and protein complement of organims, the need for the DNA sequencing of genomes will be avoided, and emphasis placed on those found to be novel.

Just as genome sequencing is not an end in itself, neither is an annotated 2-D protein reference map of an organism, nor indeed the identification of proteins in a proteome. So whilst an immediate aim of proteome projects is to screen proteins in reference maps, this will lead to expression studies and characterisation of post-translational modifications. The challenge that then needs to be addressed is the investigation of structure and function of proteins in a proteome. The magnitude of this is illustrated by the fact that over half the open reading frames identified in *S. cerevisiae* chromosome III were initially of no known function (Oliver *et al.*, 1992). Structural and functional studies will be an undertaking just as formidable as genome studies are now and proteome projects are becoming, but will lead to an unimaginably detailed understanding of how living organisms are constructed and how they operate.

Acknowledgements

MRW is the recipient of an Australian Postgraduate Research Award. AAG, MRW, IHS and KLW acknowledge assistance for proteome projects through Macquarie University Research Grants, the Australian Research Council, the Australian National Health and Medical Research Council, Beckman Instruments and GBC Scientific Equipment. DH acknowledges the financial support of a Montus Foundation Grant and the Swiss National Fund for Scientific Research (Grant # 31-33658.92). We thank colleagues who supplied work that was 'In Press' during the writing of this review.

References

ANDERSON, N.L., HOFMANN, J.P., GEMMELL, A. AND TAYLOR, J. (1984). Global approaches to quantitative analysis of gene-expression patterns observed by use of two-dimensional gel electrophoresis. *Clinical Chemistry*, 30, 2031–2036.

ANDERSON. N.L. AND ANDERSON. N.G. (1991). A two-dimensional gel database of human plasma proteins. *Electrophoresis*. 12. 853–906.

ANDERSON. N.L.. ESQUER-BLASCO. R.. HOFMANN. J.P. AND ANDERSON. N.G. (1991). A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. *Electrophoresis*. 12. 907–930.

ANDERSON. N.L.. COPPLE. D.C.. BENDELE. R.A.. PROBST. G.S.. RICHARDSON. F.C. (1992). Covalent protein modifications and gene expression changes in rodent liver following administration of methapyrilene: a study using two-dimensional electrophoresis. *Fundamental and Applied Toxicology*. 18. 570–580.

APPEL. R.D.. BAIROCH. A. AND HOCHSTRASSER. D.F. (1994). A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server. *Trends in Biochemical Sciences*. 19. 258–260.

APPEL. R.D.. HOCHSTRASSER. D.F.. FUNK. M.. VARGAS. J.R.. PELLIGRINI. C.. MULLER. A.F. AND SCHERRER. J.R. (1991). The MELANIE project: from a biopsy to automatic protein map interpretation by computer. *Electrophoresis*. 12. 722–735.

APPEL. R.D.. SANCHEZ. J-C.. BAIROCH. A.. GOLAZ. O.. MIU. M.. VARGAS. J.R. AND HOCHSTRASSER. D.F. (1993). SWISS-2DPAGE: a database of two-dimensional gel electrophoresis images. *Electrophoresis*. 14. 1323–1328.

APPEL. R.D.. SANCHEZ. J-C.. BAIROCH. A.. GOLAZ. O.. RAVIER. F.. PASQUALI. C.. HUGHES. G. AND HOCHSTRASSER. D.F. (1994). The SWISS-2DPAGE database of two-dimensional polyacrylamide gel electrophoresis. *Nucleic Acids Research*. 22. 3581–3582.

BAIROCH. A. AND BOECKMANN. B. (1994). The SWISS-PROT protein sequence databank: current status. *Nucleic Acids Research*. 22. 3578–3580.

BAKER. C.S.. CORBETT. J.M.. MAY. A.J.. YACOUB. M.H. AND DUNN. M.J. (1992). A human myocardial two-dimensional electrophoresis database: protein characterisation by microsequencing and immunoblotting. *Electrophoresis*. 13. 723–726.

BARTLET-JONES. M.. JEFFERY. W.A.. HANSEN. H.F. AND PAPPIN. D.J.C. (1994). Peptide ladder sequencing by mass spectrometry using a novel, volatile degradation reagent. *Rapid Communications in Mass Spectrometry*. 8. 737–742.

BAUER. D.. MULLER. H.. REICH. J.. RIEDEL. H.. AHRENKIEL. V.. WARTHOE. P. AND STRAUSS. M. (1993). Identification of differentially expressed mRNA species by an improved display technique (DDRT-PCR). *Nucleic Acids Research*. 21. 4272–4280.

BERNERS-LEE. T.J.. CAILIAU. R.. GROFF. J.F. AND POLLERMAN. B. (1992). *Electronic Networking: Research. Applications. and Policy*. 2. 52–58.

BIRD. A.P. (1995). Gene number, noise reduction and biological complexity. *Trends in Genetics*. 11. 94–100

BJELLQVIST. B.. EK. K.. RIGHETTI. P.G.. GIANAZZA. E.. GORG. A.. WESTERMEIER. R. AND POSTEL. W. (1982). Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *Journal of Biochemical and Biophysical Methods*. 6. 317–339.

BJELLQVIST. B.. PASQUALI. C.. RAVIER. F.. SANCHEZ. J-C. AND HOCHSTRASSER. D.F. (1993a). A nonlinear wide-range immobilized pH gradient for two-dimensional electrophoresis and its definition in a relevant pH scale. *Electrophoresis*. 14. 1357–1365.

BJELLQVIST. B.. SANCHEZ. J-C.. PASQUALI. C.. RAVIER. F.. PAQUET. N.. FRUTIGER. S.. HUGHES. G.J. AND HOCHSTRASSER. D.F. (1993b). Micropreparative 2-D electrophoresis allowing the separation of milligram amounts of proteins. *Electrophoresis*. 14. 1375–1378.

BJELLQVIST. B.. HUGHES. G.. PASQUALI. C.. PAQUET. N.. RAVIER. F.. SANCHEZ. J-C.. FRUTIGER. S. AND HOCHSTRASSER. D. (1993c). The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis*. 14. 1023–1031.

BONNER. W.M. AND LASKEY. R.A. (1974). A film detection method for tritium-labeled proteins and nucleic acids in polyacrylamide gels. *European Journal of Biochemistry*. 46. 83–88.

BOUTELL. T.. GARRELS. J.I.. FRANZA. B.R.. MONARDO. P.J. AND LATTER. G.I. (1994). REF52 on global gel navigator: an internet-accessible two-dimensional gel electrophoresis database. *Electrophoresis*. 15. 1487–1490.

BREWER. J.. GRUND. E.. HAGERLID. P.. OLSSON. I. AND LIZANA. J. (1986). In *Electrophoresis '86* (M.J. Dunn. Ed.). pp. 226–229. VCH. Weinheim.

Carr, S.A., Huddleston, M.J. and Bean, M.F. (1993). Selective identification and differentiation of *N*- and *O*-linked oligosaccharides in glycoproteins by liquid chromatography-mass spectrometry. *Protein Science*, 2, 183–196.

Campbell, K.P., MacLennan, D.H. and Jorgensen, A.O. (1983). Staining of the Ca²⁺-binding proteins, calsequesterin, calmodulin, troponin C, and S-100, with the cationic dye 'Stains-all'. *Journal of Biological Chemistry*, 258, 11267–11273.

Celis, J.E., Cruger, D., Kiil, J., Dejgarrd, K., Lauridsen, J.B., Ratz, G.P., Basse, B., Celis, A., Rasmussen, H.H., Bauw, G. and Vandekerkhove, J. (1990a). A two-dimensional gel protein database of noncultured total normal human epidermal keratinocytes: identification of proteins strongly up-regulated in psoriatic epidermis. *Electrophoresis*, 11, 242–254.

Celis, J.E., Gesser, B., Rasmussen, H.H., Madsen, P., Leffers, H., Dejgaard, K., Honore, B., Olsen, E., Ratz, G., Lauridsen, J.B., Basse, B., Mourizten, S., Hellerup, M., Andersen, A., Walbum, E., Celis, A., Bauw, G., Puype, M., Van Damme, J. and Vandekerkhove, J. (1990b). Comprehensive two-dimensional gel protein database offers a global approach to the analysis of human cells: the transformed amnion cells (AMA) master database and its link to genome DNA sequence data. *Electrophoresis*, 11, 898–1071.

Celis, J.E., Rasmussen, H.H., Olsen, E., Madsen, P., Leffers, H., Honore, B., Dejgaard, K., Gromov, P., Hoffmann, H.J., Nielsen, M., Vassilev, A., Vintermyr, O., Hao, J., Celis, A., Basse, B., Lauridsen, J., Ratz, G.P., Andersen, A.H., Walbum, E., Kjaergaard, I., Puype, M., Van Damme, J., and Vandekerkhove, J. (1993) The human keratinocyte two-dimensional database: update 1993. *Electrophoresis*, 14, 1091–1198.

Celis, J.E. and Olsen, E. (1994). A qualitative and quantitative protein database approach identified individual and groups of functionally related proteins that are differentially regulated in simian virus 40 (SV40) transformed human keratinocytes: an overview of the functional changes associated with the transformed phenotype. *Electrophoresis*, 15, 309–344.

Corbett, J.M., Dunn, M.J., Posch, A. and Gorg, A. (1994a). Positional reproducibility of protein spots in two-dimsensional polyacrylamide electrophoresis using immobilzed pH gradient isolelectric focusing in the first dimension – an interlaboratory comparison. *Electrophoresis*, 15, 1205–1211.

Corbett, J.M., Wheeler, C.H., Baker, C.S., Yacoub, M.H. and Dunn, M.J. (1994b). The human myocardial two-dimensional gel protein database: update 1994. *Electrophoresis*, 15, 1459–1465.

Cordwell, S., Wilkins, M.R., Cerpa-Poljak, A., Gooley, A.A., Duncan, M., Williams, K.L. and Humphery-Smith, I. (1995). Cross-species identification of proteins separated by two-dimensional electrophoresis using MALDI-TOF and amino acid composition. *Electrophoresis*, 15, 438–443.

Cottrell, J.S. (1994). Protein identification by peptide mass fingerprinting. *Peptide Research*, 7, 115–124.

Crimmins, D.L., McCourt, D.W., Thoma, R.S., Scott, M.G., Macke, K. and Schwartz, B.D. (1990). *In situ* chemical cleavage of proteins immobilized to glass-fiber and polyvinylidenefluoride membranes: cleavage at tryptophan residues with 2-(2'-nitrophenylsulfenyl)-3-methyl-3'-bromoindolenine to obtain internal amino acid sequence. *Analytical Biochemistry*, 187, 27–38.

Duthel, S. and Revol, A. (1993). Glycan microheterogeneity of alpha 1-antitrypsin in serum and meconium from normal and cystic fibrosis patients by crossed immunoaffinoelectrophoresis with different lectins (Con A, LCA, WGA). *Clinical and Chemical Acta*, 215, 173–187.

Eckerskorn, C., Jungblut, P., Mewes, W., Klose, J. and Lottspeich, F. (1988). Identification of mouse brain proteins after two-dimensional electrophoresis and electroblotting by microsequence analysis and amino acid composition. *Electrophoresis*, 9, 830–838.

Eckerskorn, C., Strupat, K., Karas, M., Hillenkamp, F. and Lottspeich, F. (1992).

Mass spectrometric analysis of blotted proteins after gel electrophoretic separation by matrix-assisted laser desorption/ionization. *Electrophoresis*, 13, 664–665.

ECKERSKORN, C. AND LOTTSPEICH, F. (1993). Structural characterisation of blotting membranes and the influence of membrane parameters for electroblotting and subsequent amino acid sequence analysis of proteins. *Electrophoresis*, 14, 831–838.

EK, K., BJELLQVIST, BJ. AND RIGHETTI, P.G. (1983). Preparative isoelectric focusing in immobilized pH gradients. I. General principles and methodology. *Journal of Biochemical and Biophysical Methods*, 8, 135–155.

FEY, S.J., CARLSEN, J., MOSE LARSEN, P., JENSEN, U.A., KJELDSEN, K. AND HAUNSO, S. (1994). Two-dimensional gel electrophoresis as a tool for molecular cardiology. Proceedings of the International Society for Heart Research 'XV European Section Meeting', pp 9–16

FREY, J.R., KUHN, L., KETTMAN, J.R. AND LEFKOVITS, I. (1994). The amino acid composition of 350 lymphocyte proteins. *Molecular Immunology*, 31, 1219–1231.

GARRELS, J.I. (1989). The QUEST system for quantitative analysis of two-dimensional gels. *Journal of Biological Chemistry*, 264, 5269–5282.

GARRELS, J.I. AND FRANZA, B.R. (1989) The REF52 protein database. *Journal of Biological Chemistry*, 264, 5283–5298.

GARRELS, J.I., FRANZA, B.R., CHANG, C. AND LATTER, G. (1990). Quantitative exploration of the REF52 protein database: cluster analysis reveals the major protein expression profiles in responses to growth regulation, serum stimulation, and viral transformation. *Electrophoresis*, 11, 1114–1130.

GARRELS, J.I., FUTCHER, B., KOBAYASHI, R., LATTER, I., SCHWENDER, B., VOLPE, T., WARNER, J.R. AND MCLAUGHLIN, C.S. (1994). Protein identification for a *Saccharomyces cerevisiae* protein database. *Electrophoresis*, 15, 1466–1486.

GELFI, C., BOSSI, M.L., BJELLQVIST, B. AND RIGHETTI, P.G. (1987). Isoelectric focusing in immobilized pH gradients in the pH 10–11 range. *Journal of Biochemical and Biophysical Research Methods*, 15, 41–48.

GHARAHDAGHI, F., ATHERTON, D., DEMOTT, M. AND MISCHE, S.M. (1992). Amino acid analysis of PVDF-bound proteins. in *Techniques in Protein Chemistry III* (R.H. Ageletti, Ed.), pp 249–260. Academic Press, San Diego.

GHOSH, P., OKOH, C., LIU, Q.H. AND LAKSHMAN, M.R. (1993). Effects of chronic ethanol on enzymes regulating sialylation and desialylation of transferrin in rats. *Alcoholism: Clinical and Experimental Research*, 17, 576–579.

GIOMETTI, C.S., GEMMELL, M.A., TOLLAKSEN, S.L. AND TAYLOR, J. (1991). Quantitation of human leukocyte proteins after silver staining: a study with two-dimensional electrophoresis. *Electrophoresis*, 12, 536–543

GIOMETTI, C.S., TAYLOR, J. AND TOLLAKSEN, S.L. (1992). Mouse liver protein database: a catalog of proteins detected by two-dimensional gel electrophoresis. *Electrophoresis*, 13, 970–991.

GOLAZ, O. HUGHES, G.J., FRUTIGER, S., PAQUET, N., BAIROCH, A., PASQUALI, C., SANCHEZ, J.C., TISSOT, J.D., APPEL, R.D., WALZER, C., BALANT, L. AND HOCHSTRASSER, D.F. (1993). Plasma and red blood cell protein maps: update 1993. *Electrophoresis*, 14, 1223–1231.

GOLD, M.R., YUNGWIRTH, T., SUTHERLAND, C.L., INGHAM, R.J., VIANZON, D., CHIU, R., VAN-OOSTVEEN, I., MORRISON, H.D. AND AEBERSOLD, R. (1994). Purification and identification of tyrosine-phosphorylated proteins from lymphocytes stimulated through the antigen receptor. *Electrophoresis*, 15, 441–453.

GOLDBERG, H.A., DOMENICUCCI, C., PRINGLE, G.A. AND SODEK, J. (1988). Mineral-binding proteoglycans of fetal porcine calvarial bone. *Journal of Biological Chemistry*, 263, 12092–12101.

GOOLEY, A.A., MARSHCHALEK, R. AND WILLIAMS, K.L. (1992). Size polymorphisms due to changes in the number of *O*-glycosylated tandem repeats in the *Dictyostelium discoideum* glycoprotein PsA. *Genetics*, 130, 749–756.

GORG, A., POSTEL, W. AND GUNTHER, S. (1988). The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, 9, 531–546.

GORG, A., POSTEL, W., GUNTHER, S., WESER, J., STRAHLER, J.R., HANASH, S.M., SOMERLOT, L.

and Kuick, R. (1988). Approach to stationary two-dimensional pattern: influence of focusing time and immobiline/carrier ampholyte concentrations. *Electrophoresis*, 9, 37–46.

Gravel, P., Golaz, O., Walzer, C., Hochstrasser, D.F., Turler, H., and Balant, L.P. (1994). Analysis of glycoproteins separated by two-dimensional gel electrophoresis using lectin blotting revealed by chemiluminescence. *Analytical Biochemistry*, 221, 66–71.

Gunther, S., Postel, W., Wiering, H. and Gorg, A. (1988). Acid phosphatase typing for breeding nematode-resistant tomatoes by isoelectric focusing with an ultranarrow immobilized pH gradient. *Electrophoresis*, 9, 618–620.

Hanash, S.M., Strahler, J.R., Neel, J.V., Hailat, N., Melhem, R., Keim, D., Zhu, X.X., Wagner, D., Gage, D.A. and Watson, J.T. (1991). Highly resolving two-dimensional gels for protein sequencing. *Proceedings of the National Academy of Sciences USA*, 88, 5709–5713.

Harrington, M.G., Coffman, J.A., Calzone, F.J., Hood, L.E., Britten, R.J. and Davidson, E.H. (1992). Complexity of sea urchin embryo nuclear proteins that contain basic domains. *Proceedings of the National Academy of Sciences USA*, 89, 6252–6256.

Harrington, M.G., Lee, K.H., Yun, M., Zewert, T., Bailey, J.E. and Hood, L.E. (1993). Mechanical precision in two-dimensional electrophoresis can improve spot positional reproducibility. *Applied and Theoretical Electrophoresis*, 3, 347–353.

Hellman, U., Wernstedt, C., Gonez, J. and Heldin, C-H. (1995). Improvement of an in-gel digestion for the micropreparation of internal protein fragments for amino acid sequencing. *Analytical Biochemistry*, 224, 451–455.

Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C., Grimley, C. and Watanabe, C. (1993). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proceedings of the National Academy of Sciences USA*, 90, 5011–5015.

Hess, D., Covey, T.C., Winz, R., Brownsey, R.W. and Aebersold, R. (1993). Analytical and micropreparative peptide mapping by high performance liquid chromatography/electrospray mass spectrometry of proteins purified by gel electrophoresis. *Protein Science*, 2, 1342–1351.

Hobohm, U., Holthaeve, T. and Sander, C. (1994). Amino acid analysis and protein database compositional search as a rapid and inexpensive method to identify proteins. *Analytical Biochemistry*, 222, 202–209.

Hochstrasser, D.F. and Merril, C.R. (1988). 'Catalysts' for polyacrylamide gel polymerization and detection of proteins by silver staining. *Applied and Theoretical Electrophoresis*, 1, 35–40.

Hochstrasser, D.F., Patchornik, A., and Merril, C.R. (1988). Development of polyacrylamide gels that improve the separation of proteins and their detection by silver staining. *Analytical Biochemistry*, 173, 412–423.

Hochstrasser, A.C., James, R.W., Pometta, D. and Hochstrasser, D.F. (1991a). Preparative isoelectrofocusing and high resolution two-dimensional electrophoresis for concentration and purification of proteins. *Applied and Theoretical Electrophoresis*, 1, 333–337.

Hochstrasser, D.F., Appel, R.D., Vargas, R., Perrier, R., Vurlod, J.F., Ravier, F., Pasquali, C., Funk, M., Pelligrini, C., Muller, A.F. and Scherrer, J.R. (1991b). A clinical molecular scanner: the Melanie project. *Medical Computing*, 8, 85–91.

Hochstrasser D.F., Frutiger, S., Paquet, N., Bairoch, A., Ravier, F., Pasquali, C., Sanchez, J-C., Tissot, J-D., Bjellqvist, B., Vargas, R., Appel, R.D. and Hughes, G.J. (1992). Human liver protein map: a reference database established by microsequencing and gel comparison. *Electrophoresis*, 13, 992–1001.

Holt, T.G., Chang, C., Laurent-Winter, C., Murakami, T., Davies, J.E. and Thompson, C.J. (1992). Global changes in gene expression related to antibiotic synthesis in *Streptomyces hygroscopicus*. *Molecular Microbiology*, 6, 969–980.

Honore, B., Leffers, H., Madsen, P. and Celis, J.E. (1993). Interferon-gamma up-regulates a unique set of proteins in human keratinocytes. Molecular cloning and expression of the cDNA encoding the RGD-sequence containing protein IGUP I-5111. *European Journal of Biochemistry*, 218, 421–430.

Huberty, M.C., Vath, J.E., Yu, W. and Martin, S.A. (1993). Site-specific carbohydrate identification in recombinant proteins using MALD-TOF MS. *Analytical Chemistry*. 65. 2791–2800.

HUGHES. G.J.. FRUTIGER. S.. PAQUET N.. PASQUALI. C.. SANCHEZ. J-C.. TISSOT. J.D.. BAIROCH. A.. APPEL. R.D. AND HOCHSTRASSER. D.F. (1993). Human liver protein map update 1993. *Electrophoresis*. 14. 1216–1222.

HUGHES. J.H.. MACK. K. AND HAMPARIAN. V.V. (1988). India ink staining of proteins on nylon and hydrophobic membranes. *Analytical Biochemistry*. 173. 18–25.

JAMES. P.. QUADRONI. M.. CARAFOLI. E. AND GONNET. G. (1993). Protein identification by mass profile fingerprinting. *Biochemical and Biophysical Research Communications*. 195. 58–64.

JI. H.. WHITEHEAD. R.H.. REID. G.E.. MORITZ. R.L.. WARD. L.D. AND SIMPSON. R.J. (1994) Two-dimensional electrophoretic analysis of proteins expressed by normal and cancerous human crypts: application of mass spectrometry to peptide-mass fingerprinting. *Electrophoresis*. 15. 391–405.

JOHNSON. R.S. AND WALSH. K.A. (1992). Sequence analysis of peptide mixtures by automated integration of Edman and mass spectrometric data. *Protein Science*. 1. 1083–1091.

JOHNSTON. R.F.. PICKETT, S.C. AND BARKER. D.L. (1990). Autoradiography using storage phosphor technology. *Electrophoresis*. 11. 355–360.

JUNGBLUT. P.. DZIONARA. M.. KLOSE. J. AND WITTMANN-LEIBOLD. B. (1992). Identification of tissue proteins by amino acid analysis after purification by two-dimensional electrophoresis. *Journal of Protein Chemistry*. 11. 603–612.

JUNGBLUT. P.. OTTO. A.. ZEINDL-EBERHART. E.. PLEIßNER. K-P.. KNECHT. M.. REGITZ-ZAGROSEK. V.. FLECK. E. AND WITTMANN-LEIBOLD. B. (1994). Protein composition of the human heart: the construction of a myocardial two-dimensional electrophoresis database. *Electrophoresis*. 15. 685–707.

KOHARA. Y.. AKIYAMA. K. AND ISONO. K. (1987). The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell*. 50. 495–508.

KLOSE. J. (1975). Protein mapping by combined isoelectric focusing and electrophoresis in mouse tissues. A novel approach to testing for individual point mutations in mammals. *Human Genetik*. 26. 231–243.

LATHAM. K.E.. GARRELS. J.I.. CHANG. C. AND SOLTER. D. (1991). Quantitative analysis of protein synthesis in mouse embryos I: extensive re-programming at the one- and two-cell stages. *Development*. 2. 921–932.

LATHAM. K.E.. GARRELS. J.I.. CHANG. C. AND SOLTER. D. (1992). Analysis of embryonic mouse development: construction of a high-resolution. two-dimensional gel protein database. *Applied and Theoretical Electrophoresis*. 2. 163–170.

LATHAM. K.E.. GARRELS. J.I. AND SOLTER. D. (1993) Two-dimensional analysis of protein synthesis. *Methods in Enzymology*. 255. 473–489.

LE MAIRE. M.. DESCHAMPS. S.. MOLLER. J.V.. LE CAER. J.P. AND ROSSIER. J (1993) Electrospray ionization mass spectrometry from sodium dodecyl sulfate-polyacrylamide gel electrophoresis: application to the topology of the sarcoplasmic reticulum Ca^{2+} ATPase. *Analytical Biochemistry*. 214. 50–57.

LEMKIN. P.F. AND LESTER. E.P. (1989). Database and search techniques for two-dimensional gel protein data: a comparison of paradigms for exploratory data analysis and prospects for biological modelling. *Electrophoresis*. 10. 122–140.

LEMKIN. P.F.. WU. Y. AND UPTON. K. (1993). An efficient disk-based data structure for rapid searching of quantitative two-dimensional gel databases. *Electrophoresis*. 14. 1341–1350

LI. K.W.. GERAERTS. W.P.. VAN-ELK. R. AND KOOSE. J. (1989) Quantification of proteins in the subnanogram and nanogram range: comparison of the AuroDye. FerriDye. and india ink staining methods. *Analytical Biochemistry*. 182. 44–47.

LIANG. P. AND PARDEE. A.B. (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*. 257. 967–971.

MANN. M. (1995). Sequence database searching by mass spectrometric data. In *Microcharacterisation of Proteins* (R. Kellner. F. Lottspeich. and H.E. Meyer. Eds). pp 223–245. VCH. Weinheim.

MANN, M., HOJRUP, P. AND ROEPSTORFF, P. (1993). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biological Mass Spectrometry*, 22, 338–345.

MANN, M. AND WILM, M. (1994). Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, 66, 4390–4399.

MATSUDAIRA, P. (1987). Sequence of picomole quantities of proteins electroblotted onto polyvinylidene difluoride membranes. *Journal of Biological Chemistry*, 262, 10035–10038.

MONARDO, P.J., BOUTELL, T., GARRELS, J.I. AND LATTER, G.I. (1994). A distributed system for two-dimensional gel analysis. *Computer Applications in the Biosciences*, 10, 137–143.

MORTZ, E., VORM, O., MANN, M. AND ROEPSTORFF, P. (1994). Identification of proteins in polyacrylamide gels by mass spectrometric peptide mapping combined with database search. *Biological Mass Spectrometry*, 23, 249–261.

MURTHY, L.R. AND IQBAL, K. (1991). Measurement of picomoles of phosphoamino acids by high performance liquid chromatography. *Analytical Biochemistry*, 193, 299–305.

MYRICK, J.E., LEMKIN, P.F., ROBINSON, M.K. AND UPTON, K.M. (1993). Comparison of the BioImage Visage 2000 and the GELLAB-II two-dimensional electrophoresis image analysis systems. *Applied and Theoretical Electrophoresis*, 3, 335–346.

NEIDHARDT, F.C., APPLEBY, D.B., SANKAR, P., HUTTON, M.E. AND PHILLIPS, T.A. (1989). Genomically linked cellular protein databases derived from two-dimensional polyacrylamide gel electrophoresis. *Electrophoresis*, 10, 116–122.

NIKODEM, V. AND FRESCO, J.R. (1979). Protein fingerprinting by SDS-gel electrophoresis after partial fragmentation with CNBr. *Analytical Biochemistry*, 97, 382–386.

NOKIHARA, K., MORITA, N. AND KURIKI, T. (1992). Applications of an automated apparatus for two-dimensional electrophoresis, Model TEP-1, for microsequence analysis of proteins. *Electrophoresis*, 13, 701–707.

O'FARRELL, P.H. (1975). High resolution two-dimensional electrophoresis of proteins. *Journal of Biological Chemistry*, 250, 4007–4021.

O'FARRELL, P.Z., GOODMAN, H.M. AND O'FARRELL, P.H. (1977). High resolution two-dimensional electrophoresis of basic as well as acidic proteins. *Cell*, 12, 1133–1142.

OLIVER *et al.* (1992). The complete DNA sequence of yeast chromosome III. *Nature* 357, 38–46.

OLSEN, A.D. AND MILLER, M.J. (1988). Elsie 4: quantitative computer analysis of sets of two-dimensional gel electrophoretograms. *Analytical Biochemistry*, 169, 49–70.

ORTIZ, M.L., CALERO, M., FERNANDEZ-PATRON, C., PATRON, C.F., CASTELLANOS, L. AND MENDEZ, E. (1992). Imidazole-SDS-Zn reverse staining of proteins in gels containing or not SDS and microsequence of individual unmodified electroblotted proteins. *FEBS Lettersers*, 296, 300–304.

OSTERGREN, K., ERIKSSON, G. AND BJELLQVIST, B. (1988). The influence of support material used on band sharpness in Immobiline gels. *Journal of Biochemical and Biophysical Methods*, 16, 165–170.

OU, K., WILKINS, M.R., YAN, J.X., GOOLEY, A.A., FUNG, Y., SHEUMACK, D. AND WILLIAMS, K.L. (1995). Improved high-performance liquid chromatography of amino acids derivatised with 9-fluorenylmethyl chloroformate. *Journal of Chromatography* (in press).

PACKER, N., WILKINS, M.R., GOLAZ, O., LAWSON, M., GOOLEY, A.A., HOCHSTRASSER, D.F., REDMOND, J. AND WILLIAMS, K.L. (1995). Characterisation of human plasma glycoproteins separated by two-dimensional gel electrophoresis. *Bio/Technology* (in press).

PAPPIN, D.J.C., HOJRUP, P. AND BLEASBY, A.J. (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology*, 3, 327–332.

PATTERSON, S.D. (1994). From electrophoretically separated protein to identification: strategies for sequence and mass analysis. *Analytical Biochemistry*, 221, 1–15.

PATTERSON, S.D. AND LATTER, G.I. (1993). Evaluation of storage phosphor imaging for quantitative analysis of 2-D gels using the Quest II system. *BioTechniques*, 15, 1076–1083.

PISANO, A., REDMOND, J.W., WILLIAMS, K.L. AND GOOLEY, A.A. (1993). Glycosylation sites identified by solid-phase Edman degradation: *O*-linked glycosylation motifs on human glycophorin A. *Glycobiology*, 3, 429–435.

RABILLOUD, T. (1992). A comparison between low background silver diamine and silver nitrate protein stains. *Electrophoresis*, 13, 429–439.

Rasmussen, H.H., Van Damme, J., Puype, M., Gesser, B., Celis, J.E. and Vandekerckhove, J. (1992). Microsequences of 145 proteins recorded in the two-dimensional gel protein database of normal human epidermal keratinocytes. *Electrophoresis*, 13, 960–969.

Rasmussen, H.H., Mortz, E., Mann, M., Roepstorff, P. and Celis, J.E. (1994). Identification of transformation sensitive proteins recorded in human two-dimensional gel protein databases by mass-spectrometric peptide mapping alone and in combination with microsequencing. *Electrophoresis*, 15, 406–416.

Richardson, F.C., Horn, D.M. and Anderson, N.L. (1994). Dose-responses in rat hepatic protein modification and expression following exposure to the rat hepatocarcinogen methapyrilene. *Carcinogenesis*, 15, 325–329.

Righetti, P.G. (1990). Immobilized pH gradients: theory and methodology. In *Laboratory Techniques in Biochemistry and Molecular Biology* (R.H. Burdon and P.H. van Knippenberg, Eds) Elsevier, Amsterdam.

Righetti, P.G. and Drysdale, J.W. (1973). *Annals of the New York Academy of Sciences*, 209, 163–186.

Rodriguez, L.V., Gernsten, D.M., Ramagli, L.S. and Johnston, D.A. (1993). Towards stoichiometric silver staining of proteins resolved in complex two-dimensional electrophoresis gels: real-time analysis of pattern development. *Electrophoresis*, 14, 628–637.

Rosenfeld, J., Capdevielle, J., Guillemot, J.C. and Ferrara, P. (1992). In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. *Analytical Biochemistry*, 203, 173–179.

Sanchez, J.C., Ravier, F., Pasquali, C., Frutiger, S., Paquet, N., Bjellqvist, B., Hochstrasser, D.F. and Hughes, G.J. (1992). Improving the detection of proteins after transfer to polyvinylidene difluoride membranes. *Electrophoresis*, 13, 715–717.

Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. and Petersen, G.B. (1982). Nucleotide sequence of bacteriophage λ DNA. *Journal of Molecular Biology*, 162, 729–773.

Scheele, G.J. (1975). Two-dimensional analysis of soluble proteins. *Biochemistry*, 250, 5375–5385.

Shaw, G. (1993). Rapid identification of proteins. *Proceedings of the National Academy of Sciences USA*, 90, 5138–5142.

Sibbald, P.R., Sommerfeldt, H. and Argos, P. (1991). Identification of proteins in sequence databases from amino acid composition. *Analytical Biochemistry*, 198, 330–333.

Simpson, R.J., Tsugita, A., Celis, J.E., Garrels, J.I. and Mewes, H.W. (1992). Workshop on two-dimensional gel protein databases. *Electrophoresis*, 13, 1055–1061.

Sinha, P.K., Kottgen, E., Stoffler, M-M., Gianazza, E. and Righetti, P.G. (1990). Two-dimensional maps in very acidic immobilized pH gradients. *Journal of Biochemical and Biophysical Methods*, 20, 345–352.

Smith, D.W. (1994). Introduction. In *Biocomputing: Informatics and Genome Projects* (D.W. Smith, Ed.), pp1–12. Academic Press, San Diego.

Strupat, K., Karas, M., Hillenkamp, F., Eckerskorn, C. and Lottspeich, F. (1994). Matrix-assisted laser desorption ionization mass spectrometry of proteins electroblotted after polyacrylamide gel electrophoresis. *Analytical Chemistry*, 66, 464–470.

Sutton, C.W., Pemberton, K.S., Cottrell, J.S., Corbett, J.M., Wheeler, C.H., Dunn, M.J. and Pappin, D.J. (1995). Identification of myocardial proteins from two-dimensional gels by peptide mass fingerprinting. *Electrophoresis*, 16, 308–316.

Tous, G.I., Fausnaugh, J.L., Akinyosoye, O., Lackland, H., Wintercash, P., Vitorica, F.J. and Stein, S. (1989). Amino acid analysis on polyvinylidene difluoride membranes. *Analytical Biochemistry*, 179, 50–55.

Tovey, E.R., Ford, S.A. and Baldo, B.A. (1987). Protein blotting on nitrocellulose: some important aspects of the resolution and detection of antigens in complex extracts. *Journal of Biochemical and Biophysical Methods*, 14, 1–17.

Urwin, V.E. and Jackson, P. (1993). Two-dimensional polyacrylamide gel electrophoresis of proteins labeled with the fluorophore monobromobimane prior to first-dimensional isoelectric focusing: imaging of the fluorescent protein spot patterns using a cooled charge-coupled device. *Analytical Biochemistry*, 209, 57–62.

VanBogelen, R.A., Hutton, M.E. and Neidhardt, F.C. (1990). Gene-protein database of *Escherichia coli* K-12: edition 3. *Electrophoresis*. **11**. 1131–1166.

VANBOGELEN. R.A. AND NEIDHARDT. F.C. (1991). The gene-protein database of *Escherichia coli*: edition 4. *Electrophoresis*. **12**. 955–994.

VANBOGELEN. R.A.. SANKER. P.. CLARK. R.L.. BOGAN. J.A. AND NEIDHARDT. F.C. (1992) The gene-protein database of *Escherichia coli*: edition 5. *Electrophoresis*. **13**. 1014–1054.

VANDEKERKHOVE. J.. BAUW. G.. VANCOMPERNOLLE. K.. HONORE. B. AND CELIS. J. (1990) Comparative two-dimensional gel analysis and microsequencing identifies gelsolin as one of the most prominent downregulated markers of transformed human fibroblast and epithelial cells. *Journal of Cell Biology*. **111**. 95–102.

VANFLETEREN. J.R.. RAYMACKERS. J.G.. VAN BUN. S.M. AND MEHUS. L.A. (1992). Peptide mapping and microsequencing of proteins separated by SDS-PAGE after limited *in situ* hydrolysis. *BioTechniques*. **12**. 550–557.

VORM. O. AND MANN. M. (1994) Improved mass accuracy in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of peptides. *Journal of the American Society for Mass Spectrometry*. **5**. 955–958.

VORM. O.. ROEPSTORFF. P. AND MANN. M. (1994). Improved resolution and very high sensitivity in MALDI TOF of matrix surfaces made by fast evaporation. *Analytical Chemistry*. **66**. 3281–3287.

WALLACE. A. AND SALUZ. H.P. (1992a). Ultramicrodetection of proteins in polyacrylamide gels. *Analytical Biochemistry*. **203**. 27–34.

WALLACE. A. AND SALUZ. H.P. (1992b). Beyond silver staining. *Nature*. **357**. 608–609.

WALSH. B.J.. GOOLEY. A.A.. WILLIAMS. K.L. AND BREIT. S.N. (1995). Identification of macrophage activation associated proteins by two-dimensional electrophoresis and microsequencing. *Journal of Leukocyte Biology*. **57**. 507–512.

WASINGER. V.C.. CORDWELL. S.J.. POLJAK. A.. YAN. J.X.. GOOLEY. A.A.. WILKINS. M.R.. DUNCAN. M.. HARRIS. R.. WILLIAMS. K.L. AND HUMPHERY-SMITH. I. (1995). Progress with Gene-Product Mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis*. **16**. In Press.

WEITZHANDLER. M.. KADLECEK. D.. AVDALOVIC. N.. FORTE. J.G.. CHOW. D. AND TOWNSEND. R. R. (1993). Monosaccharide and oligosaccharide analysis of proteins transferred to polyvinylidene fluoride membranes after sodium dodecyl sulfate-polyacrylamide gel electrophoresis. *Journal of Biological Chemistry*. **268**. 5121–5130.

WILKINS. M.R.. PASQUALI. C.. APPEL. R.D.. OU. K.. GOLAZ. O.. SANCHEZ. J-C.. YAN. J.X.. GOOLEY. A.A.. HUGHES. G.. HUMPHERY-SMITH. I.. WILLIAMS. K.L.. AND HOCHSTRASSER. D.F. (1995). From Proteins to Proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. Submitted.

WILKINS. M.R.. OU. K.. APPEL. R.D.. GOLAZ. O.. PASQUALI. C.. YAN. J.X.. FARNSWORTH. V.. CARTIER. P.. HOCHSTRASSER. D.F.. WILLIAMS. K.L. AND GOOLEY. A.A. (1996) Rapid protein identification using N-terminal sequence tagging and amino acid analysis (submitted)

WIRTH. P.J.. LUO. L.D.. FUJIMOTO. Y.. BISGAARD. H.C. AND OLSEN. A.D. (1991). The rat liver epithelial (RLE) cell protein database. *Electrophoresis*. **12**. 931–954.

WIRTH. P.J.. LUO. L.D.. BENJAMIN. T.. HOANG. T.N.. OLSEN. A.D. AND PARMALEE. D.C. (1993) The rat liver epithelial (RLE) cell nuclear protein database. *Electrophoresis*. **14**. 1199–1215.

WU. Y.. LEMKIN. P.F. AND UPTON. K. (1993). A fast spot segmentation algorithm for two-dimensional gel electrophoresis analysis. *Electrophoresis*. **14**. 1351–1356.

YAMAGUCHI. K. AND ASAKAWA. H. (1988). Preparation of colloidal gold for staining proteins electrotransferred onto nitrocellulose membranes. *Analytical Biochemistry*. **172**. 104–107.

YAMASHITA. K.. IDEO. H.. OHKURA. T.. FUKUSHIMA. K.. YUASA. I.. OHNO. K. AND TAKESHITA. K. (1993) Sugar chains of serum transferrin from patients with carbohydrate deficient glycoprotein syndrome. Evidence of asparagine-N-linked oligosaccharide transfer deficiency. *Journal of Biological Chemistry*. **268**. 5783–5789.

YATES. J.R. III. SPEICHER. S.. GRIFFIN. P.R. AND HUNKAPILLER. T. (1993). Peptide mass maps: a highly informative approach to protein identification. *Analytical Biochemistry*. **214**. 397–408.

3

The

R.M. D

Therm
Hamilt
Wales
Privat

Introd

Proteas
dealing
type. T
structur
since th
by Was
1930) c
possibl
mechan
subtilis
1986; I
comme
proteas
the chai
been in
regulat
apparer
Kalisz.
Barrett
Give
surprisi
attracti
subject
stabilit
enzym
Sherod

• Corres

Biotechn
0264-872

D

Human cellular protein patterns and their link to genome DNA sequence data: usefulness of two-dimensional gel electrophoresis and microsequencing

JULIO E. CELIS,[*,1] HANNE H. RASMUSSEN,[*] HENRIK LEFFERS,[*] PEDER MADSEN,[*] BENT HONORÉ,[*] BORBALA GESSER,[*] KURT DEJGAARD,[*] JOËL VANDEKERCKHOVE[†]

[*] *Institute of Medical Biochemistry and Human Genome Research Centre, Aarhus University, DK-8000 Aarhus, Denmark and* [†] *Laboratorium voor Fysiologische Chemie, Rijksuniversiteit Gent, Belgium*

ABSTRACT **Analysis of cellular protein patterns by computer-aided 2-dimensional gel electrophoresis together with recent advances in protein sequence analysis have made possible the establishment of comprehensive 2-dimensional gel protein databases that may link protein and DNA information and that offer a global approach to the study of the cell. Using the integrated approach offered by 2-dimensional gel protein databases it is now possible to reveal phenotype specific protein (or proteins), to microsequence them, to search for homology with previously identified proteins, to clone the cDNAs, to assign partial protein sequence to genes for which the full DNA sequence and the chromosome location is known, and to study the regulatory properties and function of groups of proteins that are coordinately expressed in a given biological process. Human 2-dimensional gel protein databases are becoming increasingly important in view of the concerted effort to map and sequence the entire genome.—Celis, J. E.; Rasmussen, H. H.; Leffers, H.; Madsen, P.; Honoré, B.; Gesser, B.; Dejgaard, K.; Vandekerckhove, J. Human cellular protein patterns and their link to genome DNA sequence data: usefulness of two-dimensional gel electrophoresis and microsequencing. *FASEB J.* 5: 2200-2208; 1991.**

Key Words: human protein patterns · 2-dimensional gel protein databases · gene expression · microsequencing · cDNA cloning · linking protein and DNA information · genome mapping and sequencing

PROTEINS SYNTHESIZED FROM information contained in the DNA orchestrate most cellular functions. The total number of proteins synthesized by a typical human cell is unknown although current estimates range from 3000 to 6000. Of these, as many as 70% may perform household functions and are expected to be shared by all cell types irrespective of their origin. There are many different cell types in the human body with perhaps 30,000 to 50,000 proteins expressed in the organism as a whole judged from the fact that about 3% of the haploid genome correspond to genes. Today only a small fraction of the total set of proteins has been identified, and little is known about the protein patterns of individual cell types or their variation under physiological and abnormal conditions.

For the past 15 years, high resolution 2-dimensional gel electrophoresis has been the technique of choice to determine the protein composition of a given cell type and for monitoring changes in gene activity through quantitative and qualitative analysis of the thousands of proteins that orchestrate various cellular functions (refs 1-6 and references therein). The technique originally described by O'Farrell (separates proteins in terms of their isoelectric point (pI) an.. molecular weight. Usually one chooses a condition of interest and the cell reveals the global protein behavioral response as all detected proteins can be analyzed both qualitatively and quantitatively in relation to each other. At present, most available 2-dimensional gel techniques (regular gel format) can resolve between 1000 and 2000 proteins from a given mammalian cell type, a number that corresponds to about 2 million base pairs of coded DNA. Less abundant proteins can be detected by analyzing partiall purified cellular fractions.

Two-dimensional gel ectrophoresis has been widely applied to analysis of cellular protein patterns from bacteria to mammalian cells (refs 1-6, and references therein). In spite of much work, however, information gathered from these studies has not reached the scientific community in its fullness because of lack of standardized gel systems and the lack of means for storing and communicating protein information. Only recently, because of the development of appropriate computer software (7-13), has it been possible to scan gels, assign numbers to individual proteins, and store the wealth of information in quantitative and qualitative comprehensive 2-dimensional gel protein databases (4, 14-23), i.e., those containing information about the various properties (physical, chemical, biological, biochemical, physiological, genetic, immunological, architectural, etc.) of all the proteins that can be detected in a given cell type. Such integrated 2-dimensional gel protein databases offer an easy and standardized medium in which to store and communicate protein information and provide a unique framework in which to focus a multidisciplinary approach to study the cell. Once a protein is identified in the database, all of the information accumulated can be easily retrieved and made available to the researcher. In the long run, protein databases are expected to foster a wide variety of biological information that may be instrumental to researchers working in many areas of biology—among others, cancer and oncogene studies, differentiation, development, drug development and testing, genetic variation, and diagnosis of genetic and clinical diseases (Fig. 1).

The approach using systematic 2-dimensional gel protein analysis has recently gained a new dimension with the advent of techniques to microsequence major proteins recorded

[1] To whom correspondence should be addressed, at: Institute of Medical Biochemistry and Human Genome Research Centre, Ole Worms Alle, Bldg. 170, University Park, DK-8000 Aarhus C, Denmark.

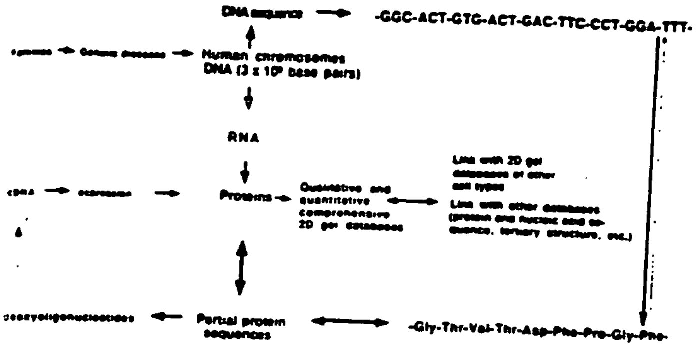

DNA sequence
-GGC-ACT-GTG-ACT-GAC-TTC-CCT-GGA-TTT-
Human chromosomes
DNA (3 x 10^9 base pairs)
RNA
cDNA
Proteins
Qualitative and quantitative comprehensive 2D gel databases
Link with 2D gel databases of other cell types
Link with other databases (protein and nucleic acid sequence, tertiary structure, etc.)
Partial protein sequences
-Gly-Thr-Val-Thr-Asp-Phe-Pro-Gly-Phe-

Figure 1. Interface between partial protein sequence databases. comprehensive 2-dimensional gel databases. and the human genome sequencing project. Appropriate software is required to compare protein and DNA sequences. In general. although the inference of a protein's sequence from the DNA sequence (thick arrow) is direct and unambiguous. the DNA sequence can only be inferred approximately from the protein sequence (thin arrow) and cloning of the gene requires either a cDNA or the requisite group of oligonucleotide probes deduced from the partial amino acid sequence. Modified from ref 6.

in the databases (refs 24–42 and references therein). Partial protein sequences can be used to search for protein identity as well as to prepare specific DNA probes for cloning as-yet-uncharacterized proteins (Fig. 1). As these sequences can be stored in the database (see for example Fig. 2H), they offer a unique opportunity to link information on proteins with the existing or forthcoming DNA sequence data on the human genome (Fig. 1) (20. 36. 39).

Using the integrated approach offered by comprehensive 2-dimensional gel databases (Fig. 1), it will be possible to identify phenotype-specific proteins; microsequence them and store the information in the database: search for homology with previously characterized proteins; clone the cDNAs. assign partial protein sequences to genes for which the full DNA sequence and the chromosome location are known. and study the regulatory properties and function of groups of proteins (pathways. organelles, etc.) that are coordinately expressed in a given biological process. Comprehensive 2-dimensional gel protein databases will depict an integrated picture of the expression levels and properties of the thousands of protein components of organelles. pathways. and cytoskeletal systems in both physiological and abnormal conditions and are expected to lead to identification of new regulatory networks in different cell types and organisms. In the future. 2-dimensional gel protein databases may be linked to each other as well as to national and international specialized databanks on nucleic acid and protein sequences. protein structures. NMR experimental data. complex carbohydrates. etc.

A few 2-dimensional gel protein databases that are accessible in a computer form have been published in extenso: these correspond to the protein-gene database of *Escherichia coli* K-12 developed by Neidhardt and colleagues (14. 23), the rat REF 52 database established by Garrels and co-workers at Cold Spring Harbor (18. 22). and a few human databases (transformed amnion cells [15. 20]. normal embryonal lung MRC-5 fibroblasts [17. 21]. keratinocytes [19] and peripheral blood mononuclear cells [15]) developed in Aarhus. Given space limitations and to keep this review in focus. we will concentrate on the computerized analysis of human cellular 2-dimensional gel patterns. and in particular on the steps involved in establishing comprehensive 2-dimensional gel databases that can link protein and DNA information.

MAKING AND MANAGING A COMPREHENSIVE 2-DIMENSIONAL GEL DATABASE OF HUMAN CELLULAR PROTEINS

The first step in making a comprehensive 2-dimensional gel protein database is to prepare a synthetic image (digital form of the gel image) of the gel (fluorogram. Coomassie blue or silver stained gel) to be used as a standard or master reference. This can be done with laser scanners. charge couple device (CCD)[2] array scanners. television cameras. rotating drum scanners. and multiwire chambers (13). Computerized analysis systems for spot detection. quantitation. pattern matching, and data handling (access and retrieval of information. database making) have been described in the literature (ELSIE [43]. GELLAB [11]. HERMeS [44]. MELANIE [10]. QUEST (9), and TYCHO [8]) and some are available commercially (PDQUEST. Protein Database Inc.. Huntington. N.Y.; KEPLER, Large Scale Biology. Rockville. Md.; Visage, BioImage Corporation. Ann Arbor. Mich.; Gemini. Joyce Loebl, Gateshead: Microscan 1000. Technology Resources Inc., Nashville, Tenn. and MasterScan, Billerica. Mass.). Unfortunately, most of these systems are incompatible with one another and their advantages and disadvantages have been discussed by Miller (13).

In our work station in Aarhus. fluorograms are scanned with a Molecular Dynamics laser scanner and the data are analyzed using the PDQUEST II software (Protein Databases Inc.) (12) running on a spark station computer 4100 FC-8-P3 from SUN Microsystems, Inc. The scanner measures intensity in the range of 0–2.0 absorbance. A typical scan of a 17 × 17 cm fluorogram takes about 2 min. Steps in image analysis include: initial smoothing, background substraction, final smoothing, spot detection, and fitting of ideal Gaussian distribution to spot centers. Spot intensity is calculated as the integration of a fitted Gaussian. If calibration strips containing individual segments of a known amount of radioactivity are used, it is possible to merge multiple exposures of the sample image into a single data image of greater dynamic range. Once the synthetic image is created it can be stored on disk and displayed directly on the monitor. Functions that can be used to edit the images include: cancel (for example, to erase scratches that may have been interpreted as spots by the computer; cancel streaks or low dpm spots), combine (sometimes a spot may be resolved into several closely packed spots), restore, uncombine, and add spot to the gel. The process is time consuming—about 1-1/2 day per image. Edited standard images can be matched to other synthetic images. Figure 2*A* shows a portion of a standard synthetic image (IEF) of a fluorogram of [^{35}S]methionine labeled cellular proteins from human AMA cells (master database) (20). Images can be displayed either in black and white (resembling the original fluorograms) or in color (other images in Fig. 2), depending on the need. As shown in Fig. 2*B*, each polypeptide is assigned a number by the computer, which facilitates the entry and retrieval of qualitative and quantitative information for any given spot in the gel (20). The standard image can be matched automatically by the computer to other standard or reference gels (Fig. 2*C*, matching of AMA cellular proteins [left] to MRC-5 proteins [right]) provided a few landmark spots are given manually as reference (indicated with a + in Fig. 2*C*) to initiate the process.

[2] Abbreviations: CCD. charge couple device: PCNA. proliferating cell nuclear antigen: HPLC. high performance liquid chromatography.

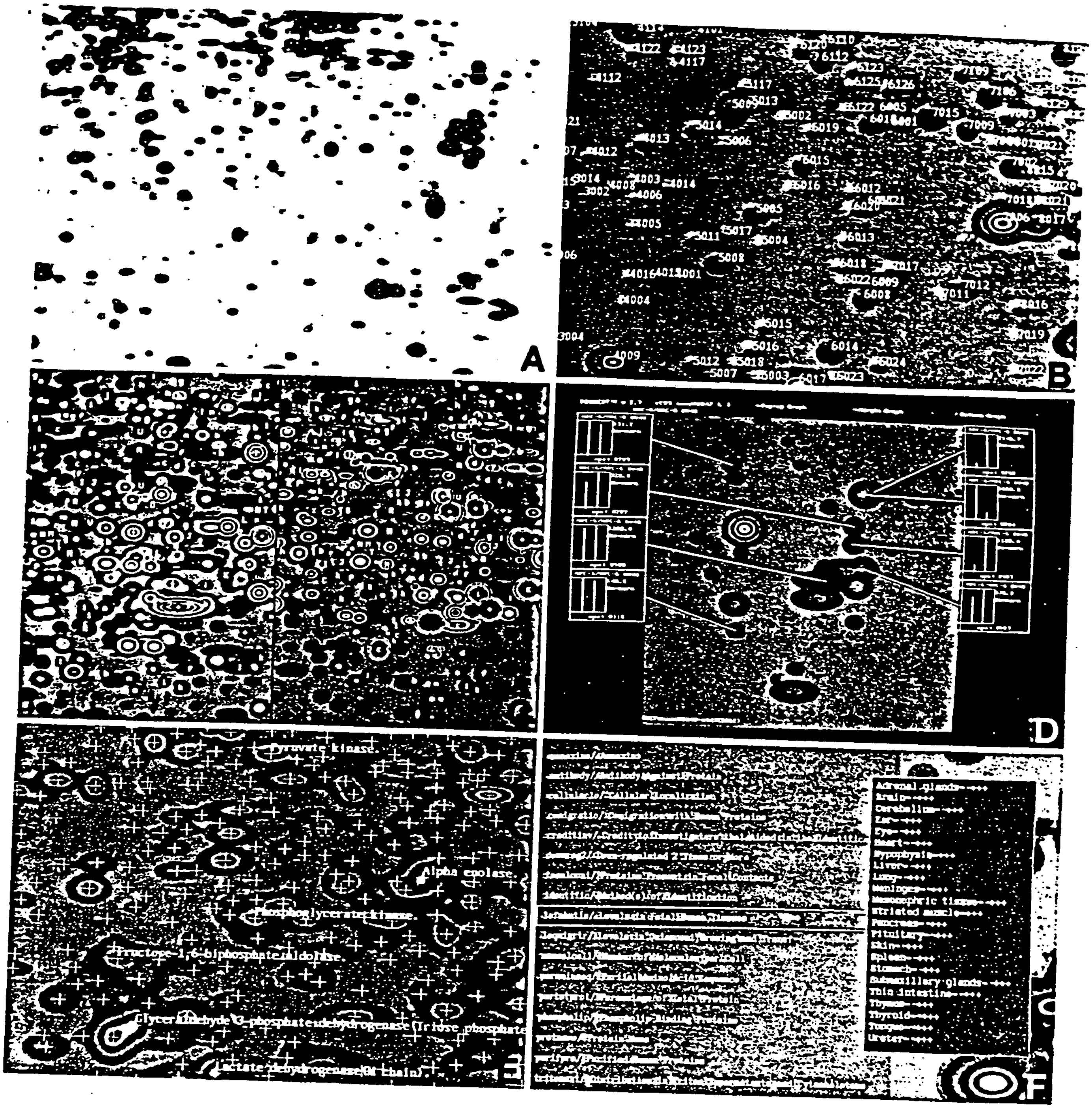
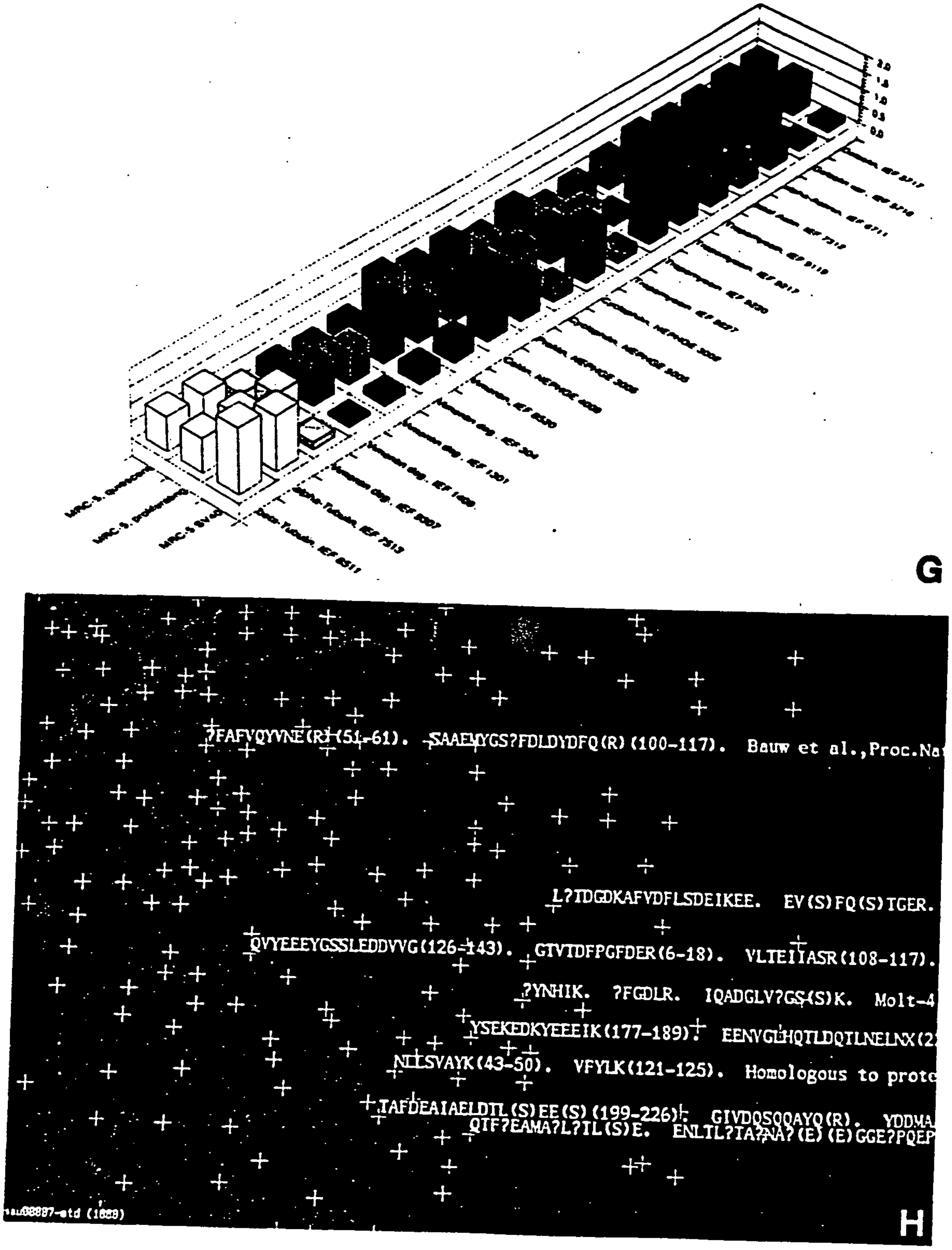


Figure 2. *A*) Synthetic image of a fraction of an IEF gel of the master image of AMA cellular proteins. *B*) As in *A* but showing numbers assigned to each spot. *C*) Comparison of AMA (left) and normal human embryonal lung MRC-5 fibroblasts (right) IEF proteins patterns. Matched proteins are indicated by a + or by the same letters in both gels. Once a protein is matched, information contained in the various categories available in the master AMA database can be transferred. *D*) Synthetic image of a fraction of an IEF fluorogram of [^{35}S]methionine labeled proteins from normal human MRC-5 fibroblasts. The histograms show levels of synthesis of a few proteins in MRC-5 (left bar) and SV40 transformed MRC-5 (right bar) fibroblasts. *E*) Polypeptides that contain information under the category glycolytic pathway. *F*) The function peruse annotation for spot allows the operator to inquire about categories and information available for a given protein. *G*) Relative abundance of cytoskeletal and cytoskeletal-related proteins in quiescent, proliferating, and SV40-transformed MRC-5 fibroblasts. *H*) Polypeptides that contain information under the category partial amino acid sequences.



The automatic matching process that has been described in detail by Garrels et al. (12) takes about 5 min. Matched proteins are indicated with the same letters in both gels (Fig. 2C). The usefulness of this function is emphasized by the fact that data accumulated on common household proteins can be easily transferred to any other human cellular cell type whose 2-dimensional gel cellular protein pattern is matched to our standard AMA 2-dimensional gel protein image. Alternatively, if the standard gel is part of a matchset (set of gels in a given experiment) it can be used as a linker gel to compare, for example, the quantitative values of a given protein throughout the experiment (see Fig. 2D; levels of some proteins in normal and SV40 transformed human MRC-5 fibroblasts) or with other standard images in different sets of cross-matched experiments (18, 22).

Once a standard map of a given protein sample is made, one can enter qualitative annotations to make a reference database. Our master 2-dimensional gel database of transformed human amnion cell (AMA) proteins (20) lists 3430 polypeptides of which 2592 correspond to cellular components, having pI's ranging from 4 to 13 and molecular weights between 8.5 and 230 kDa. The most abundant proteins in the database correspond to total actin (3.87% of total protein; about 90 million molecules per cell) while the lesser abundant of the recorded polypeptides are present in the vicinity of 5000 molecules per cell. Some annotation categories we are using to establish the master AMA database include: *1)* protein identification (comigration with purified proteins, 2-dimensional immunoblotting, microsequencing); *2)* amounts (total amounts and levels of synthesis); *3)* subcellular localization (nuclear, cytoskeletal, membrane, membrane receptors, specific organelles, etc.); *4)* antibodies; *5)* posttranslational modifications (phosphorylation, glycosylation, methylation etc.); *6)* microsequencing; *7)* cell cycle specificity (specific variations in levels of synthesis and amount); *8)* regulatory behavior (effect of hormones, growth factors, heat shock, etc.) *9)* rate of synthesis in normal and transformed cells (proliferation sensitive proteins, cell cycle specific proteins, oncogenes, components of the pathway (or pathways) that control cell proliferation); *10)* function (mainly from comigration with proteins of known function); *11)* sets of proteins that are coordinately regulated (hierarchy of controls, differential gene expression in various cells, etc.); *12)* cDNAs (cloned cDNAs); *13)* proteins that are specific to a given disease (systematic comparison of protein patterns of fibroblast proteins from healthy and diseased individuals); *14)* expression and exploitation of transfected cDNAs; *15)* pathways (metabolic, others); *16)* gene localization (genetic and physical); *17)* effect of microinjected antibody on patterns of protein synthesis; and *18)* secreted proteins.

Information entered for any spot in a given annotation category can be easily retrieved by asking the computer to display the information on the color screen. For example, Fig. 2*E* shows a synthetic image of a NEPHGE gel (master AMA database) displaying the information contained under the entry glycolytic pathway. Alternatively, one can use the function peruse annotations for spot to directly ask the computer to list all the entries available for a particular protein. By clicking the mouse in a given entry (in this case, presence in fetal human tissues) it is possible to take a quick look at the information in that particular entry (Fig. 2*F*).

A major obstacle encountered in building comprehensive 2-dimensional gel protein databases is identifying the large number of proteins separated by this technology. In our databases (20, 21), known proteins are identified by one or a combination of the following procedures: *1)* comigration with known proteins, *2)* 2-dimensional gel immunoblotting using specific antibodies, and *3)* microsequencing of Coomassie Brillant Blue stained human proteins recovered from dried 2-dimensional gels (see next section). Protein identification by means of microsequencing may be difficult, as individual protein members of families with short peptide differences may escape detection. In the gene-protein database of *E. coli* K-12 (14, 23), another major 2-dimensional gel database available at present, proteins are being identified by a wider range of tests that include comigration with purified proteins; genetic criterion (deletion, insertion, frameshift, nonsense, missense, regulatory), plasmid-bearing strains and in vitro synthesis of protein; selective labeling (methylation, phosphorylation); peptide map similarity; and physiological criterion and selective derivatization.

So far we have received nearly 550 antibodies from laboratories all over the world and these are being systematically tested by 2-dimensional gel immunoblotting for antigen determination. Similarly, purified proteins and organelles provided by several laboratories have greatly aided identification of unknown proteins (20, 21). We routinely request antibodies and protein samples and promise the donors to make available all the information we may have accumulated on that particular protein. For example, Table 1 lists entries available for Lipocortin V (IEF SSP 8216), also known as annexin V, VAC-α, endonexin II, renocortin, chromobindin-5', anticoagulant protein, PAP-I, γ-calcimedin, IBC, calphobindin, and anchorin CII.

As mentioned previously, one distinct advantage of 2-dimensional gel electrophoresis is the possibility of studying quantitative variations in cellular protein patterns that may lead to identification of groups of proteins that are expressed coordinately during a given biological process. Quantitation, however, is not an easy task as reflected by the lack of published data on global cellular protein patterns. We believe this is partly due to difficulties in obtaining sets of gels that are suitable for computer analysis (streaking, material remaining at the origin, etc.) as well as to limitations (laborious editing time, need of calibration strips to merge images, limited dynamic range, etc.) in the computer analysis systems available at the moment. Perhaps the most advanced quantitative studies published so far using computer analysis have been carried out by Garrels and coworkers (18, 22). In particular, these investigators have established a quantitative rat protein database (18, 22) designed to study growth control (proliferation, growth inhibitors, and stimulation) and transformation in well-defined groups of cell lines obtained by transformation of rat REF52 cells with SV40, adenovirus, and the Kirsten murine sarcoma virus. These studies have revealed clusters of proteins induced or repressed during growth to confluence as well as groups of transformation-sensitive proteins that respond in a differential fashion to transformation by DNA and RNA viruses. A most interesting feature of this quantitative database is the discovery of a group of coregulated proteins that show similar expression patterns as the cell cycle-regulated DNA replication protein known as proliferating cell nuclear antigen (PCNA)/cyclin (45).

In our human databases, most quantitations have been carried out by estimating the radioactivity contained in the polypeptides by direct counting of the gel pieces in a scintillation counter (20, 21). Up to 700 proteins can be cut out through appropriate exposed films in a period of time comparable to that required for editing a synthetic image. Manual quantitation of this large number of spots is difficult without the assistance of a master reference image and a numbering system that can be used to identify the spots. Using this approach, we have recorded quantitative changes in the relative abundance of 592 [^{35}S]methionine-labeled proteins synthesized by quiescent, proliferating, and SV40 transformed human embryonic lung MRC-5 fibroblasts (21). Some data concerning cytoskeletal and cytoskeletal-related proteins are presented in Fig. 2G. Our studies as well as those of Garrels and co-workers (18, 22) may in the long run help define patterns of gene expression that are characteristic of the transformed state.

OTHER 2-DIMENSIONAL GEL PROTEIN DATABASES

As mentioned previously there are other 2-dimensional gel databases available in computer form that have been pub-

TABLE 1. *Some entries for lipocortin V in the human AMA 2-dimensional gel protein database*

Entries for lipocortin V (IEF SSP 8216)	Information entered
1. Protein name	Lipocortin V, renocortin, chromobindin-5', endonexin I, anticoagulant protein, PAP-I, VAC-α, 35-γ-calcimedin, IBC, calphobindin I, anchorin CII, annexin V
2. Percentage of total protein	0.110% (about 2,800,000 molecules per cell)
3. Apparent molecular weight (*mr*)	33.3 kDa
4. Isoelectric point (pl)	4.76
5. Method (or methods) of identification	Microsequencing, 2-dimensional immunoblotting, Comigration
6. Credit to investigators that aided in identification	G. Bauw, J. Vandekerckhove, and colleagues, Rijksuniversiteit Gent; B. Pepinsky, BIOGEN, Cambridge; N.G. Ahn, University of Washington
7. Antibody against protein	Polyclonal (rabbit, antibody no. 20), B. Pepinsky, BIOGEN, Cambridge
8. Comigration with human proteins	Lipocortin V,N.G. Ahn, Howard Hughes Medical Institute, Washington University
9. Cellular localization	Subcortical membrane
10. Calcium/phospholipid-dependent membrane proteins	Lipocortin V
11. Function	Regulation of various aspects of inflammation, immune response, blood coagulation and differentiation
12. Partial amino acid sequence	GTVTDFPGFDER (7–18), VLTEIIASR (109–117), QVYEEEYGSSLEDDVVG (127–143), ?GTDEEKFITIFGT(R) (187–201)
13. cDNA sequence	Known, R. Blake et al., *J. Biol. Chem.* 263, 10799–10811; 1988 (pl = 4.76 from translated sequence)

14. Levels in fetal human tissues	Adrenal glands = + + +; brain = + + +;
	cerebellum = + + +; ear = + + +; eye = + + +;
	heart = + + +; hypophysis = + + +; liver = + + +;
	lung = + + +; meninges = + + +;
	mesonephric tissue = + + +;
	striated muscle = + + +; pancreas = + + +;
	skin = + + +; spleen = + + +; stomach = + + +;
	submandibular gland = + + +;
	small intestine = + + +; thymus = + + +;
	thyroid gland = + + +; tongue = + + +;
	ureter = + + +

15. Levels in quiescent, proliferating, and transformed MRC-5 fibroblasts	Q (quiescent) = 1.1; P (proliferating) = 1.0; T (SV40 transformed) = 0.3
16. Distribution in Triton supernatant and cytoskeletons	Mainly supernatant

lished in extenso: these correspond to the *E. coli* K-12 protein-gene database (14, 23) and to the rat REF52 database (18, 22).

The *E. coli* K-12 cellular protein-gene database is perhaps the most complete of all databases reported so far and eventually it should trace each protein back to its structural gene. Information contained in this database includes: gene/protein name (protein name, EC number, gene name); 2-dimensional gel spot designations (x-y coordinates from reference gels, alphanumeric designation); genetic information (linkage map location, physical map location, Genebank code, sequence reference, location on Kohara clones); biochemical information (molecular weight, pI, number of residues of each amino acid, mole percent of each amino acid, total number of amino acids in a polypeptide), and regulatory information (cellular level of protein in different media and different temperature, member of regulon, member of stimulon). Major advances of this database are envisaged in the future in view of the eminent sequencing of the whole *E. coli* genome as well as the development of improved methods to express cloned genes.

The rat REF52 2-dimensional gel protein database lists about 1600 proteins that have been recorded using the QUEST analysis system (18, 22). Included in this quantitative database are *1*) protein names (cytoskeletal and heat shock proteins as well as various nuclear, mitochondrial, and cytoplasmic proteins), *2*) annotations (subcellular localization, modification, recognition by specific antibodies, coprecipitation, NH_2-terminal sequence, cross-reference to protein sequence information and references to the literature), *3*) protein sets (cytoskeletal proteins, phosphoproteins, sets of proteins with PCNA/cyclin-like properties, etc.) and *4*) general quantitative data (protein synthesis during growth of normal REF52 cells to confluence and quiescence, and after restimulation of growth-inhibited cells).

In addition to the 2-dimensional gel databases mentioned so far there are several smaller cellular databases being established in human (normal human diploid fibroblasts, lymphocytes, leukocytes, leukemic cells) mouse (NIH/3T3 cells, T lymphocytes), *Aplysia*, yeast (*Saccharomyces cerevisae*), plants (wheat, barley, sorghum), and *Euglena*. Databases of tissue protein, (brain, whole mouse, liver) and body fluid proteins (plasma proteins, cerebrospinal fluid, urine, and milk) are being established in several laboratories. The reader is directed to the review by Celis et al. (4) for details and references concerning these databases.

MICROSEQUENCING HAS ADDED A NEW DIMENSION TO COMPREHENSIVE 2-DIMENSIONAL GEL DATABASES: A DIRECT LINK BETWEEN PROTEINS AND GENES

The development of highly sensitive amino acid gas-phase or liquid-phase sequenators (24), together with the establishment of efficient protein and peptide sample preparation methods, has opened the possibility to perform a systematic sequence analysis of proteins resolved by 2-dimensional gel electrophoresis. Indeed, generated pieces of protein sequences can be used to search for protein identity (comparison with available sequences stored in databanks) as well as for preparing specific DNA probes for cloning of as yet uncharacterized proteins (Fig. 1). In addition, partial protein sequences can be stored in 2-dimensional gel databases (for example, see Fig. 2*H*) and offer a unique link between proteins and genes (Fig. 1).

In the early 1970s gel electrophoresis was used to purify proteins for sequencing purposes (reviewed by Weber and Osborn in ref 25). Proteins were recovered by diffusion and sequenced by the manual dansyl-Edman degradation at the nanomole level. This technique was further refined by using electro-elution to recover proteins and by miniaturizing the system (26). This method has been used extensively, but showed increasing drawbacks (low yields, protein samples contaminated by free amino acids, and NH_2-terminal blocking) as the amounts of handled protein gradually became smaller (e.g., at the 10 picomol level).

Most of the problems referred to above have been minimized with the introduction of protein-electroblotting procedures (27-32). When proteins are blotted on chemically inert membranes, it is possible to sequence the immobilized proteins directly without additional manipulations. Thus, depending on the amount of bound protein and its nature, this direct sequencing procedure generally yields NH_2-terminal sequences containing 10-40 residues. As such, this technique was used to identify, by their NH_2-terminal sequences, differentially expressed major proteins from total cellular extracts separated on 2-dimensional gels. A major difficulty encountered in this procedure is the occurrence of frequent artefactual blockage of the proteins. Several studies suggest that this phenomenon is mainly due to reaction with contaminants (particularly unpolymerized acrylamide present in the gel) and to a high dilution of the protein (low concentration of the protein per unit membrane surface). In addition to this primarily technical problem, many proteins are blocked in vivo by acylation or by a pyrrolidon carboxylic acid cap.

The problem of partial or complete NH_2-terminal blockage can be circumvented by generating internal amino acid sequences. This is achieved by fragmenting the protein present in the gel (gel in situ cleavage) or by cleaving it while bound to the membrane (membrane in situ cleavage) (33-35). In both cases, proteins are either cleaved in a restricted way (e.g., by limited enzymatic digestion or by using restriction chemical cleavage conditions) or fragmented into smaller peptides.

Of the different combinations examined, we had good results by using exhaustive proteolytic digestion on membrane-immobilized proteins. This method has been described for Ponceau red-stained proteins on nitrocellulose blots (34), for Amido-black-stained Immobilon-bound proteins, and for fluorescamine-detected proteins on glass fiber membranes (35). The proteases used (trypsin, chymotrypsin, or pepsin) cleave at multiple sites, generating small peptides that elute from the blot into the digestion buffer from which they are purified by reversed-phase high performance liquid chromatography (HPLC) before being sequenced individually. Although each of these manipulations could be expected to result in a reduced yield of final sequence information, we were surprised that the peptides could be sequenced with high efficiency. In our hands, this approach could be routinely applied to gel-purified proteins available in amounts ranging from 5 to 10 µg, and often yielded sequence information covering more than 30% of the total protein. As membrane-immobilized proteins are not homogeneously digested, but rather show protease sensitivity next to resistant regions, the number of peptides generated is much lower than expected from the number of potential cleavage sites. Consequently, HPLC peptide chromatograms are less complex and most peptides can be recovered in pure form.

As only limited amounts of a protein mixture can be loaded on a 2-dimensional gel, proteins of interest are often obtained in yields insufficient for the currently available sequencing technology. More material can be obtained by enriching for a certain subcellular fraction (purified cell organelles) or by exploiting affinity (dyes, metals, drugs, etc) or hydrophobic properties of proteins before gel analysis. All of the sequencing results accumulated so far in the human protein database (20) (a few are shown in Fig. 2*H*) have been obtained from analysis of protein spots collected from 2-dimensional gels that had been stained with Coomassie blue according to standard procedures and dried for storage. Proteins are recovered from the collected gel pieces by a protein-elution-concentration device, combined with gel electrophoresis and electroblotting. Details of this technique have been reported in a previous communication (42) and a brief outline is given below.

Combined gel pieces are allowed to swell in gel sample buffer (a total volume of 1.5 ml). The gel pieces combined with the supernatant are then collected into a large slot made in a new gel. The slot is further filled with Sephadex G-10 equilibrated in gel sample buffer. During consecutive gel electrophoresis, most of the electrical current passes on the side of the slot instead of passing through the slot. This results in both a vertical stacking and horizontal contraction of the protein band. With this device the protein is efficiently eluted from the gel pieces and concentrated from a large volume into a narrow spot. The highly concentrated (about 5 mm^2) protein spot is then electroblotted on PVDF-membranes, stained with Amido black, and in situ digested with trypsin. The peptides generated during digestion elute from the membrane into the supernatant, and can be separated by narrow bore reversed-phase HPLC and collected individually for sequence analysis.

Using this and previous procedures (37, 39, 42), we have so far analyzed 70 protein spots collected from 2-dimensional gels (20, and unpublished observations) (see for example Fig. 2*H*). The sequence information amounts to 2100 allocated residues corresponding to an average of 30 residues per protein spot. So far we have made cDNAs of many of the unknown proteins that have been microsequenced, and a substantial number has been cloned and sequenced. All available information indicates that it may be possible to obtain partial sequence information from most of the proteins that can be visualized by Coomassie Brillant Blue staining.

Partial protein sequences are stored in the database as displayed in Fig. 2*H*. and it should be possible in the near future to interface this information with forthcoming DNA sequence data from the human genome project. In the long run. as the human genome sequences become available it will be possible to assign partial protein sequences to genes for which the full DNA sequence and chromosomal location are known (Fig. 1).

SUMMARY

The studies presented in this brief review are intended to demonstrate the usefulness of computer-aided 2-dimensional gel electrophoresis and microsequencing to analyze cellular protein patterns. and to link protein and DNA information. As more information is gathered worldwide. comprehensive databases will depict an integrated picture of the expression levels and properties of the thousands of proteins that orchestrate most cellular functions.

Clearly, databases allow easy access to a large body of data and provide an efficient medium to communicate standardized protein information. In the future. databases will foster a wide variety of biological information that can be used to support collaborative research projects in basic and applied biology as well as in clinical research (2, 5, 46). Once a protein is identified in a particular database all the information gathered on it can be made available to the scientist. However, many problems must be solved before protein databases become of general use to the scientific community. A most urgent one is to promote standardization of the gel running conditions so that data produced in a given laboratory may be used worldwide. Surprisingly, the gel running technology as it stands today is still a craftmanship art.

Finally, comprehensive, computerized databases of proteins, together with recently developed techniques to microsequence proteins. offer a new dimension to the study of genome organization and function (Fig. 1). In particular, human protein databases may become increasingly important in view of the concerted effort to map and sequence the entire human genome. This formidable task is expected to dominate biological research in the next decades.

We would like to thank S. Himmelstrup Jørgensen for typing the manuscript and O. Sønderskov for photography. Work in the authors' laboratories was supported by grants from the Danish Biotechnology Programme. the Danish Cancer Foundation. and the Commission of the European Communities.

REFERENCES

1. O'Farrell. P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* 250. 4007–4021
2. Special Issue: Two-dimensional gel electrophoresis. *Clin. Chem.* 28. 1982
3. Celis. J. E.. and Bravo. R.. eds. (1984) *Two-Dimensional Gel Electrophoreis of Proteins: Methods and Applications.* Academic. New York
4. Celis. J. E.. Madsen. P.. Gesser. B.. Kwee. S.. Nielsen. H. V.. Rasmussen. H. H., Honoré. B.; Leffers. H.. Ratz. G. P., Basse, B.. Lauridsen, J. B.. and Celis. A. (1989) Protein databases derived from the analysis of two-dimensional gels. In *Advances in Electrophoresis* (Chrambach. C., ed) VCH, Weinheim. Germany
5. Special Issue: Two-dimensional gel electrophoresis in cell biology. (Celis. J. E.. ed) *Electrophoresis* 11. 1990
6. Celis. J. E., Honoré. B.. Bauw. G.. and Vandekerckhove. J. (1990) Comprehensive computerized 2D gel protein databases offer a global approach to the study of the mammalian cell. *BioEssays* 12. 93–98
7. Garrels. J. I. (1983) Two-dimensional gel electrophoresis and computer analysis of proteins synthesized by cloned cell lines *Methods Enzymol.* 100. 411–423
8. Anderson. N. L., Hofmann. J. P., Gemmel. A.. and Taylor. S. (1984) Global approaches to the quantitative analysis of gene-expression patterns observed by two-dimensional gel electrophoresis. *Clin. Chem.* 30. 2031-2036
9. Garrels. J. I., Farrar. J. T.. and Burwell. C. B. (1984) The Quest system for computer-analyzed two-dimensional electrophoresis of proteins in *Two-Dimensional Gel Electrophoresis of Proteins. Methods and Applications* (Celis. J. E.. and Bravo. R.. eds) pp. 37–91. Academic, New York
10. Vincens. P., and Tarroux. P. (1988) Two-dimensional electrophoresis computerized processing. *Int. J. Biochem.* 20. 499–509
11. Appel. R., Hochstrasser. D.. Roch. C.. Funk. M., Muller. A. F.. and Pellegrini. C. (1988) Automatic classification of two-dimensional gel electrophoresis pictures by heuristic clustering analysis: a step toward machine learning. *Electrophoresis* 9. 136–142
12. Lemkin. P. F., and Lester. E. P. (1989) Database and search techniques for two-dimensional gel protein data: a comparison of paradigms for exploratory data analysis and prospects for biological modeling. *Electrophoresis* 10. 122–140
13. Miller, M. J. (1989) Computer-assisted analysis of two-dimensional gel electrophoretograms. *Adv. Electrophoresis* 3. 182–217
14. Phillips, T. D., Vaughn. V., Bloch, P. L., and Neidhardt. F. C. (1987) In *Eschericia coli and Salmonella typhimurium: Cellular and Molecular Biology, Gene-Protein Index of Escherichia coli K-12*, 2 ed. (Neidhardt. F. C., Ingraham. J. I., Low. K. B., Magasanik. B., Schaechter, M., and Umbarger. H. E. ed) pp. 919–966. American Society for Microbiology, Washington, D.C.
15. Celis, J. E., Ratz, G. P., Celis, A., Madsen, P., Gesser, B., Kwee, S., Madsen, P. S., Nielsen. H. V., Yde, H., Lauridsen, J. B., and Basse, B. (1988) Towards establishing comprehensive databases of cellular proteins from transformed human epithelial amnion cells (AMA) and normal peripheral blood mononuclear cells. *Leukemia* 9, 561–601
16. Special Issue: Protein databases in two-dimensional electrophoresis. (Celis, J. E.. ed) *Electrophoresis* 2, 1989
17. Celis, J. E., Ratz. G. P., Madsen. P.. Gesser. B:, Lauridsen. J. B., Brogaard-Hansen, K. P., Kwee, S., Rasmussen. H. H.. Nielsen, H. V., Crüger. D., Basse, B., Leffers, H., Honoré. B., Møller, O., and Celis. A. (1989) Computerized. comprehensive databases of cellular and secreted proteins from normal human embryonic lung MRC-5 fibroblasts: identification of transformation and/or proliferation sensitive proteins. *Electrophoresis* 10. 76–115
18. Garrels, J. I., and Franza, B. R. (1989) The REF52 protein database. Methods of database construction and analysis using the Quest system and characterizations of protein patterns from proliferating and quiescent REF52 cells. *J. Biol. Chem.* 264, 5283–5298
19. Celis. J. E., Crüger. D.. Kiil, J., Dejgaard. K., Lauridsen, J. B., Ratz, G. P., Basse. B., Celis. A., Rasmussen, H. H., Bauw, G., and Vandekerckhove, J. (1990) A two-dimensional gel protein database of noncultured total normal human epidermal keratinocytes: identification of proteins strongly up-regulated in psoriatic epidermis. *Electrophoresis* 11, 242–254
20. Celis, J. E., Gesser, B., Rasmussen. H. H., Madsen, P., Leffers, H., Dejgaard. K., Honoré, B., Olsen, E., Ratz, G., Lauridsen, J. B., Basse, B., Mouritzen, S., Hellerup, M., Andersen, A., Walbum, E., Celis, A., Bauw, G., Puype, M., Van Damme, J., and Vandekerckhove, J. (1990) Comprehensive two-dimensional gel protein databases offer a global approach to the analysis of human cells: the transformed amnion cells (AMA) master database and its link to genome DNA sequence data. *Electrophoresis* 12, 989–1071

21. Celis, J. E., Dejgaard, K., Madsen, P., Leffers, H., Gesser, B., Honoré, B., Rasmussen, H. H., Olsen, E., Lauridsen, J. B., Ratz, G., Mouritzen, S., Hellerup, M., Andersen, A., Walbum, E., Celis, A., Bauw, G., Puype, M., Van Damme, J., and Vandekerckhove, J. (1990) The MRC-5 human embryonal lung fibroblast two-dimensional gel cellular protein database: quantitative identification of polypeptides whose relative abundance differs between quiescent, proliferating and SV40 transformed cells. *Electrophoresis* 12, 1072-1113

22. Garrels, J. I., Franza, B. R., Chang, C., and Latter, G. (1990) Quantitative exploration of the REF52 protein database: cluster analysis reveals the major protein expression profiles in responses to growth regulation, serum stimulation, and viral transformation. *Electrophoresis* 12, 1114-1130

23. Van Bogelen, R. A., Hutton, M. E., and Neidhardt, F. C. (1990) Gene-protein database of *Escherichia coli* K-12, 3rd ed. *Electrophoresis* 12, 1131-1166

24. Hewick, R. M., Hunkapiller, M. W., Hood, L. E., and Dreyer, W. J. (1981) A gas-liquid solid phase peptide and protein sequenator. *J. Biol. Chem.* 256, 7990-7997

25. Weber, K., and Osborn, M. (1985) In *The Proteins and Sodium Dodecyl Sulfate: Molecular Weight Determination on Polyacrylamide Gels and Related Procedures* (Neurath, H. et al., eds) Vol. 1, pp. 179-223, Academic, New York

26. Hunkapiller, M. W., Lujan, E., Ostrander, F., and Hood, L. E. (1983) Isolation of microgram quantities of proteins from polyacrylamide gels for amino acid sequence analysis. *Methods Enzymol.* 91, 227-236

27. Vandekerckhove, J., Bauw, G., Puype, M., Van Damme, J., and Van Montagu, M. (1985) Protein-blotting on polybrene-coated glass-fiber sheets. *Eur. J. Biochem.* 152, 9-19

28. Aebersold, R. H., Teplow, D. B., Hood, L. E., and Kent, S. B. H. (1986) Electroblotting onto activated glass. *J. Biol. Chem.* 261, 4229-4238

29. Bauw, G., De Loose, M., Inzé, D., Van Montagu, M., and Vandekerckhove, J. (1987) Alterations in the phenotype of plant cells studied by NH_2-terminal amino acid-sequence analysis of proteins electroblotted from two-dimensional gel-separated total extracts. *Proc. Natl. Acad. Sci. USA* 84, 4806-4810

30. Matsudaira, P. (1987) Sequence from picomole quantities of proteins electroblotted onto polyvinylidene difluoride membranes. *J. Biol. Chem.* 262, 10035-10038

31. Eckerskorn, C., Mewes, W., Goretzki, H., and Lottspeich, F. (1985) A new siliconized-glass fiber as support for protein-chemical analysis of electroblotted proteins. *Eur. J. Biochem.* 176, 509-519

32. Moos, M., Jr., Nguyen, N. Y., and Liu, T.-Y. (1988) Reproducible high yield sequencing of proteins electrophoretically separated and transferred to an inert support. *J. Biol. Chem.* 263, 6005-6008

33. Kennedy, T. E., Gawinowicz, M. A., Barzilai, A., Kandel, E. R., and Sweatt, J. D. (1988) Sequencing of proteins from two-dimensional gels by using in situ digestion and transfer of peptides to polyvinylidene difluoride membranes: application to protein associated with sensitization in *Aplysia*. *Proc. Natl. Acad. Sci. USA* 85, 7008-7012

34. Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E., and Kent, S. B. H. (1987) Internal amino acid sequence analysis of protein separated by one- or two-dimensional gel electrophoresis after in situ protease digestion on nitrocellulose. *Proc. Natl. Acad. Sci. USA* 84, 6970-6974

35. Bauw, G.; Van Den Bulcke, M., Van Damme, J., Puype, M., Van Montagu, M., and Vandekerckhove, J. (1988) Protein electroblotting on polybase-coated glassfiber and polyvinylidine difluoride membranes: an evaluation. *J. Prot. Chem.* 7, 194-196

36. Celis, J. E., Ratz, G. P., Madsen, P., Gesser, B., Lauridsen, J. B., Leffers, H., Rasmussen, H. H., Nielsen, H. V., Crüger, D., Basse, B., Honoré, B., Möller, O., Celis, A., Vandekerckhove, J., Bauw, G., Van Damme, J., Puype, M., and Van Den Bulcke, M. (1989) Comprehensive, human cellular protein databases and their implication for the study of genome organization and function. *FEBS Lett.* 244, 247-254

37. Bauw, G., Van Damme, J., Puype, M., Vandekerckhove, J., Gesser, B., Lauridsen, J. B., Ratz, G. P., and Celis, J. E. (1989) Protein-electroblotting and -microsequencing strategies in generating protein databases from two-dimensional gels. *Proc. Natl. Acad. Sci. USA* 86, 7701-7705

38. Aebersold, R., and Leavitt, J. (1990) Sequence analysis of proteins separated by polyacrylamide gel electrophoresis. Towards an integrated protein database. *Electrophoresis* 11, 517-527

39. Bauw, G., Rasmussen, H. H., Van Den Bulcke, M., Van Damme, J., Puype, M., Gesser, B., Celis, J. E., and Vandekerckhove, J. (1990) Two-dimensional gel electrophoresis, protein electroblotting and microsequencing: a direct link between proteins and genes. *Electrophoresis* 11, 528-536

40. Tempst, P., Link, A. J., Riviere, L. R., Fleming, M., and Elicone, C. (1990) Internal sequence analysis of protein separated on polyacrylamide gels at the submicrogram level: improved methods, applications and gene cloning strategies. *Electrophoresis* 11, 537-553

41. Eckerskorn, C., and Lottspeich, F. (1990) Combination of two-dimensional gel electrophoresis with microsequence and amino acid composition analysis: improvement of speed and sensitivity in protein characterization. *Electrophoresis* 11, 554-561

42. Rasmussen, H. H., Van Damme, J., Bauw, G., Puype, M., Gesser, B., Celis, J. E., and Vandekerckhove, J. (1991) In *Methods in Protein Sequence Analysis* (Jörnvall, H., and Höög, J. O., eds) pp. 103-114, Eighth International Conference on Methods in Protein Sequence Analysis, Birkhäuser Verlag, Boston

43. Olson, A. D., and Miller, M. J. (1988) Elsie 4: quantitative computer analysis of sets of two-dimensional gel electrophoretograms. *Anal. Biochem.* 169, 49-70

44. Vincens, P., Paris, N., Pujol, J. L., Gaboriaud, C., Rabilloud, T., Pennetier, J., Matherat, P., and Tarroux, P. (1986) HERMeS: a second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part I: Data acquisition. *Electrophoresis* 7, 347-356

45. Celis, J. E., Madsen, P., Celis, A., Nielsen, H. V., and Gesser, B. (1987) Cyclin (PCNA, auxiliary protein of DNA polymerase-δ) is a central component of the pathway(s) leading to DNA replication and cell division. *FEBS Lett.* 220, 1-7

46. Anderson, N. G., and Anderson, N. L. (1982) The human protein index. *Clin. Chem.* 28, 739-748

F

Electrophoresis 1993. 14. 1045–1053

Bo Franzén[1]
Stig Linder[3]
Ken Okuzawa[2]
Harabumi Kato[2]
Gert Auer[1]

[1]Division of Tumor Pathology, Department of Pathology, Division of Experimental Oncology, Karolinska Hospital and Institute, Stockholm Sweden
[2]Tokyo Medical College, Department of Surgery, Tokyo
[3]Division of Experimental Oncology, Karolinska Hospital and Institute, Stockholm

Nonenzymatic extraction of cells from clinical tumor material for analysis of gene expression by two-dimensional polyacrylamide gel electrophoresis

We have compared different methods of preparation of malignant cells for two-dimensional electrophoresis (2-DE). We found all methods using fresh tissue to be superior compared to methods using frozen tissue. Our results indicate that nonenzymatic methods of preparation of tumor cells, including fine needle aspiration, scraping and squeezing, have advantages over methods using enzymatic extraction of cells. Nonenzymatic methods are rapid, appear to reduce loss of high molecular protein species, and alleviate the necessity of separating viable and nonviable cells by Percoll gradient centrifugation. Using these techniques, high-quality 2-DE maps were derived from tumors of the lung and breast. In the resulting polypeptide patterns, heat shock proteins, non-muscle tropomyosins and intermediate filament were identified. We conclude that nonenzymatic extraction of malignant cells from fresh tumor tissue improves the possibilities that these techniques may be useful in clinical diagnosis.

1 Introduction

Tumors may develop by a number of different mechanisms in any given cell type. At the time of diagnosis, tumors will have progressed along different pathways to various stages of malignancy. To provide a basis for individual therapy it is of importance to examine specific properties of the tumor cell population in each patient. A large number of different markers have been described in order to increase the diagnostic accuracy. It is likely that a combination of serveral markers is needed in the future in order to reflect different properties of the tumor. One important method for the resolution of a large number of potential markers is two-dimensional electrophoresis (2-DE). Extensive efforts are being made in identifying various polypeptides separated by 2-DE and to characterize how the expression of these polypeptides is affected by the response to cellular transformation and various culture conditions [1,2]. It would be of value to transfer this information to 2-DE separations of polypeptides from tumor tissue samples. However, one prerequisite is that the quality of the 2-DE gels from tumor samples is comparable in quality with 2-DE gels from samples of cultured cells.

Frozen tumor tissues are commonly used for various biochemical assessments. However, if such samples are analyzed by 2-D polyacrylamide gel electrophoresis (PAGE), the polypeptide patterns are obscured by contamination of serum- and connective tissue proteins. Such nontumor-cell-related variations represent serious problems in the interpretation and inter-patient comparison of 2-DE patterns [3]. 2-DE patterns of cells prepared from fresh tumor material were analyzed after enzymatic extraction of tumor cells [4, 5] or after culturing tumor fragments in medium containing radioactive amino acids [6]. These procedures may, however, lead to alterations in the gene expression/polypeptide patterns. We are only aware of one study where nonenzymatic extraction of cells from fresh tumor tissue (prostate cancer) was used to prepare samples for 2-D PAGE [4]. We have examined enzymatic extraction and various nonenzymatic preparation techniques, including fine needle aspiration, for the preparation of cells from fresh tumor tissues. We describe nonenzymatic extraction procedures that are rapid, lead to high-quality 2-DE patterns, and that alleviate the necessity to purify tumor cell populations from dead cells.

2 Materials and methods

2.1 Cell cultures and samples used for spot identification

A rat embryonal fibroblast cell line, WT2 (a kind gift from Dr. J. I. Garrels and Dr. S. Pattersson) was used for the identification of a number of heat shock and structural proteins. Human normal diploid lung fibroblasts, WI38, human epithelial breast carcinoma cells, MDA-231 and MCF-7 were purchased from ATCC and grown as recommended. Polypeptides prepared from a leukemia type pre-B-ALL were separated by 2-DE. The 2-DE map was then analyzed by Dr. S. M. Hanash (University of Michigan, Ann Arbor, USA).

2.2 Tumor tissues samples

In this study, 2-DE maps from seven tumors were used as representative illustrations: two adenocarcinoma of the lung (LA, and LB, mucinous, both cases intermediate grade of differentiation), one sqamous carcinoma of the lung (LS), one carcinoid-like breast cancer (BC), one microfolliculary adenoma (highly differentiated) of the thyroid (TA), one highly differentiated hyperneph-

Correspondence: Dr. Bo Franzén, Division of Tumor Pathology, Department of Pathology, L1:01, Karolinska Hospital and Institute, 10401 Stockholm 60, Sweden

Abbreviations: **2-DE**, Two-dimensional polyacrylamide gel electrophoresis; **IEF**, isoelectric focusing; **LDH**, lactate dehydrogenase; **NP-40**, Nonidet P-40; **PBS**, phosphate buffered saline; **PCNA**, proliferating cell nuclear antigen; **PIH**, protease inhibitors; **PMSF**, phenylmethyl sulfonyl fluoride; **SDS**, sodium dodecyl sulfate; **WW**, wet weight

0173-0835/93/1010-1045 $5.00+.25/0

roma, a tumor of the kidney (KH), and finally one case of poorly differentiated corpus carcinoma (CP).

2.3 Preparation of cultured cells

The cell monolayers were washed twice in phosphate buffered saline (PBS) and then scraped off in ice-cold PBS including protease inhibitors (PIH), phenylmethylsulfonyl fluoride (PMSF) 0.2 mM and 0.83 mM benzamidine pelleted at 660 × g, 3 min (+4°C) and washed one time before final centrifugation at 2700 × g, 5 min. The wet weight of the cell pellet was recorded and the cells were stored at −80°C until further processing.

2.4 Preparation of tumor tissue samples

2.4.1 General remarks

Macroscopically representative and non-necrotic tumor tissues were selected within 20 min after resection. Parallel samples were routinely prepared for cytology. The samples were processed as rapidly as possible on ice or at +4°C and in the presence of PIH. Cells were stained with DiffQuick (Baxter) and usually examined at three different occasions during the preparation procedure: (i) cytology sample, (ii) extracted cells and (iii) cells after percoll gradient centrifugation.

2.4.2 Specimen acquisition

The strategy of sample preparation is shown in Fig. 1. Tumor tissue cell samples were usually obtained by fine needle aspiration (NA) using a 0.7 mm needle. The syringe was filled with 1–2 mL of ice-cold culture medium/PIH. We found that if a tumor appeared to be very fibrous it is difficult to extract enough cells for 2-DE analysis. In these cases, two alternative techniques were examined. (i) The tumor was cut in the middle and the fresh surface scraped (SC) by a scalpel. The cell-rich material was then transferred to ice-cold culture medium (L15 with 5% fetal calf serum)/PIH. (ii) A part of the tumor sample was placed in culture medium on ice for further processing at the laboratory in the following way: the material was cut into very small fragments on a pre-cooled dissection plate and transferred to a small glass chamber with a 0.7 mm metal net 5 mm above the bottom of the chamber. Medium /PIH was added to cover the sample (8 mL) which was gently squeezed (SQ) towards the net in order to release and wash out cells. NA and SC were also compared with an enzymatic extraction (EE) procedure described previously [5]: Briefly, thin slices of tissue were incubated with collagenase (1 mg/mL) and elastase (2 mg/mL) in medium for 1 h at 37°C. Extracted cells from every sample were then subjected to percoll gradient centrifugation (Section 3.2.3).

2.4.3 Separation of cells by Percoll gradient centrifugation

The cell suspension was filtered through two nylon mesh filters, (i) 250 μm and (ii) 100 μm and then centrifuged at 660 × g for 3 min. The cell pellet was resuspended carefully in medium, using a syringe and loaded onto a two-step discontinuous Percoll/PBS gradient, 20.4 (density = 1.03 g/mL) and 54.7% (density = 1.07 g/mL), and centrifuged at 1000 × g for 15 min. In this system, dead cells stay on the top, viable cells sediment to the interphase and erythrocytes sediment to the bottom. The viability of cells in the top fraction and interphase was checked by the trypan blue exclusion test. The interphase cell layer (> 90% viability) was collected and washed one time in a large volume PBS/PIH (centrifuged at 800 × g for 3 min). Finally, the cells were resuspended in 1.4 mL PBS and pelleted at 2700 × g for 5 min. The wet weight (WW) was recorded and the pellet was then stored at −80°C.

2.4.4 Final preparation of cells for 2-D PAGE analysis

From this point, cultured cell samples were treated in the same way as tumor cell samples: Each cell pellet was thawed on ice and resuspended in 1.89 μL mQ water per mg WW (= 1.89 × WW) μL. The suspension was frozen and thawed 4–5 × to break the cells [7]. A volume of (0.089 × WW) μL 10% sodium dodecyl sulfate (SDS), including 33.3% mercaptoethanol, was mixed with the sample and incubated 5 min on ice with (0.329 × WW) μL of a solution of DNase I (0.144 mg/mL 20 mM Tris-HCl with 2 mM $CACl_2$ × $2H_2O$, pH 8.8) and RNase A (0.0718 mg/mL Tris) [8,9]. The sample was frozen and lyophilized. Sample buffer [10] including

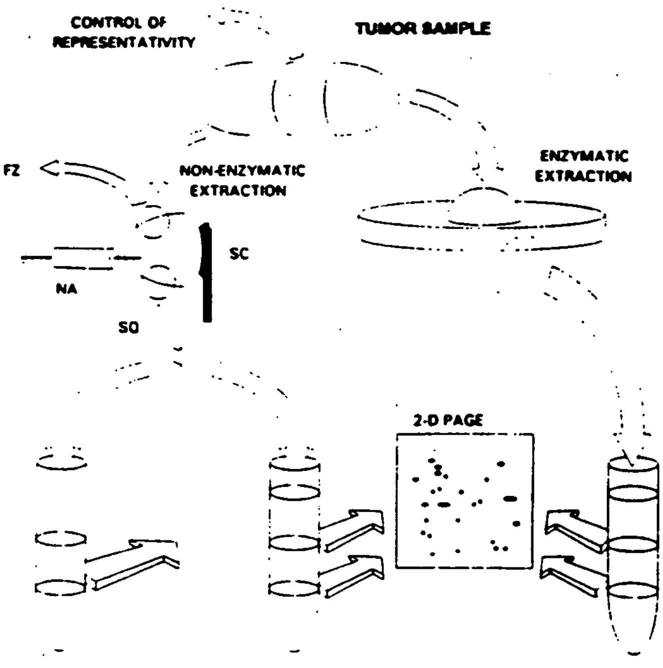


Figure 1. Experimental flow chart showing main steps of the preparation procedures. The abbreviations used for nonenzymatic extraction procedures are: FZ: frozen sample preparation; NA, needle aspiration; SC, scraped; and SQ, squeezed sample. Extracted cells are then loaded as a suspension (top volume of each tube) onto either 1.07 g/mL Percoll (left), or a discontinuous Percoll gradient from the nonenzymatic extraction (middle), or from enzymatic extraction (right). Cellular top- and interphase fractions are then used for 2-DE. For details see Section 2.

PMSF (0.2 mM, EDTA (1.0 mM), 0.5% Nonidet P-40 (NP-40), and 3-[3-cholamido propyl)-dimethylammonio]-1-propane sulfonate (CHAPS: 25 mM) was added carefully, mixed for 2.5 h and centrifuged for 15 min at 10000 rpm to remove any insoluble material. Duplicate or triplicate samples were taken for protein determination [11]. Samples were stored at −80°C prior to isoelectric focusing (IEF).

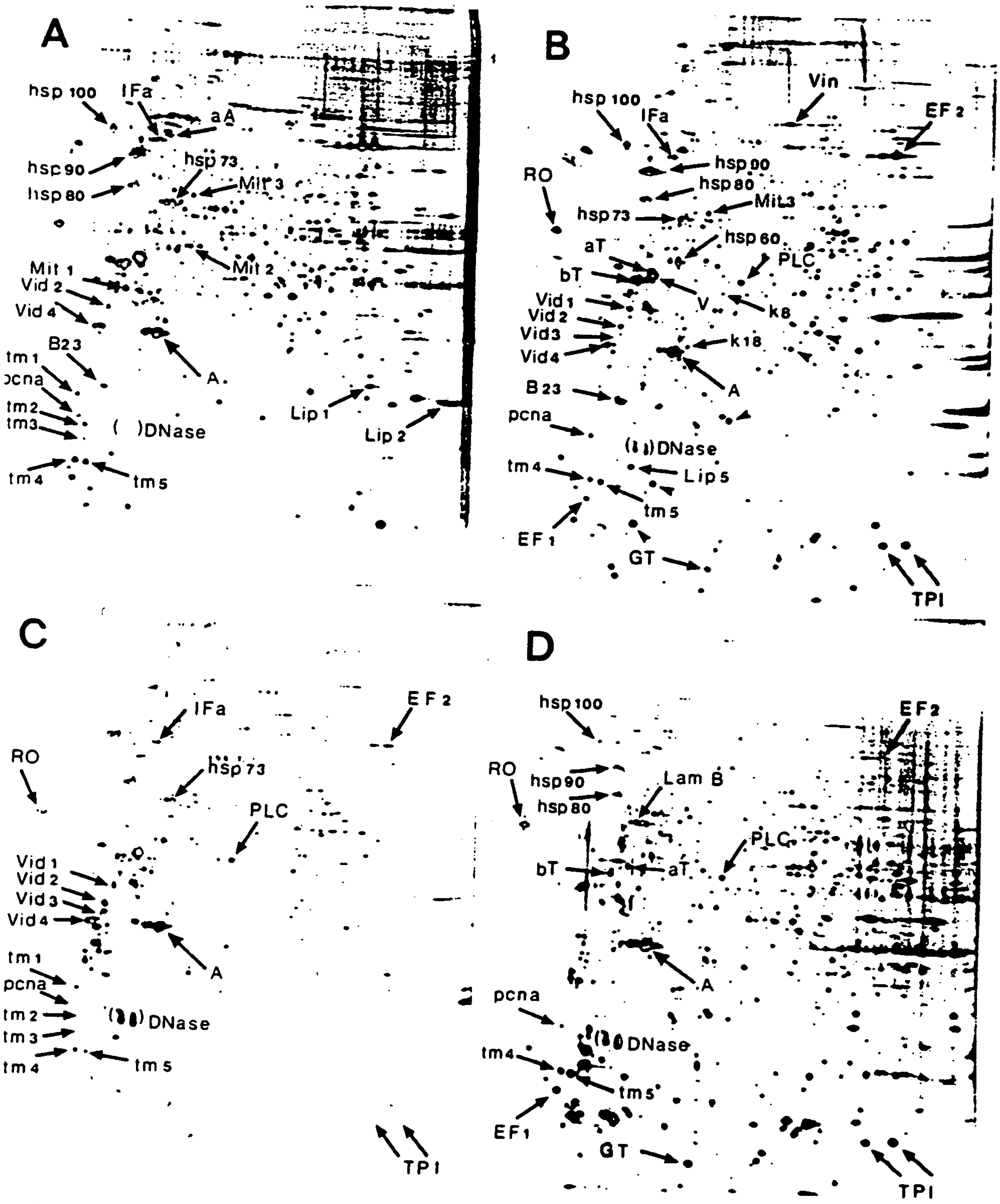


Figure 2. 2-DE analysis of samples from three cell lines and one leukemia used for the identification of polypeptides: (A) WT2; (B) MDA-231, arrowheads mark some low molecular weight cytosolic polypeptides; (C) WI38 and (D) pre B-All. The abbreviations for identified spots are explained in Table 1.

2.4.5 Preparation of frozen tumor tissue

The technique has been described previously [3.12]. Briefly, the sample is moarted frozen to a fine powder, homogenized, lyophilized and solubilized in sample buffer.

2.4.6 Control of representativity

The tumors were examined routinely by experienced pathologists and smears or imprints from the samples were also assessed for cytometric DNA content by microspectrophotometry.

2.5 2-D PAGE

2-D PAGE was performed as described [8.10] except for the following details. The glass tubes for IEF, 1.2 × 200 mm, contained 2.0% Resolyte, pH 4–8 (BDH) and were cast to a height of 180 mm. A stock solution of acrylamide (Serva) and *N,N'*-methylenebisacrylamide (16.7:1 for IEF and 37.5:1 for the second dimension) was deionized by mixing with 5% w/v Duolite MB 5313 mixed-resin ion exchanger (BDH) for 30 min, filtered (with a 0.22 µm nitrocellulose filter) and stored at −70°C. *N,N'*-Methylenebisacrylamide, *N,N,N'*,N'-tetramethylethylenediamine (TEMED) and ammonium persulfate were purchased from Bio-Rad. IEF tubes were prefocused at 200 V in 60 min. To each tube a sample corresponding to 20–40 µg protein was applied and focused for 14.5 h at 800 V and finally 1.0 h at 1000 V using a Protean II cell (Bio-Rad) and Model 1000/500 Power Supply (Bio-Rad). The tube gels were finally extruded into 1.25 mL equilibration buffer, containing 60 mM Tris, pH 6.8 (2% SDS, 100 mM dithiothreitol and 10% glycerol), frozen on dry ice and stored at −70°C. The second dimension (1.0 × 180 × 90 mm) of the acrylamide concentration was 10% T, and the gel contained 376 mM Tris, pH 8.8, and 0.1% SDS. IEF gels were applied on top of the slab gel, sealed with 0.5% agarose containing electrophoresis running buffer (60 mM Tris-base, 0.2 M glycine and 0.1% SDS) and electrophoresed with 10–11 mA per gel (constant current) at +10°C. Six gels were run together in a Protean II xi 2-D Multi-Cell (Bio-Rad). Proteins were visualized by silver staining and photographed with the acidic side to the left [13.14].

2.6 Identification of polypeptides

Vimentin and vimentin-derived polypeptides were identified by extraction of an MDA-231 cell lysate with 0.6 M KCl/0.5% NP-40 [15]. Tropomyosins were exctracted from MDA-231 and WI38 cell lysates [16], and cytokeratins were extracted from MDA-231 and MCF-7 cell lysates [17]. The patterns were compared with published maps [19–21]. Proliferating cell nuclear antigen (PCNA) was identified by immunoblotting (PC10 mAB, Dakopatt) using a semidry system (Multiphor II Nova Blot, Pharmacia-LKB Biotechnology AB) and enhanced chemoluminescence (ECL) detection (Amersham).

3 Results

3.1 2–DE of samples prepared from normal and tumorigenic cultured cells

The object of this study was to develop methods for preparation of 2-DE maps from human tumor tissue which have the same high resolution as those obtained from cultured cells. Shown in Fig. 2 are high resolution 2-DE gels prepared from cultured cells and one leukemia: SV40 transformed embryonal rat fibroblasts WT2 (Fig. 2a); human MDA-231 breast carcinoma cells (Fig. 2b); human WI38 fibroblasts (Fig. 2c) and human pre B-ALL

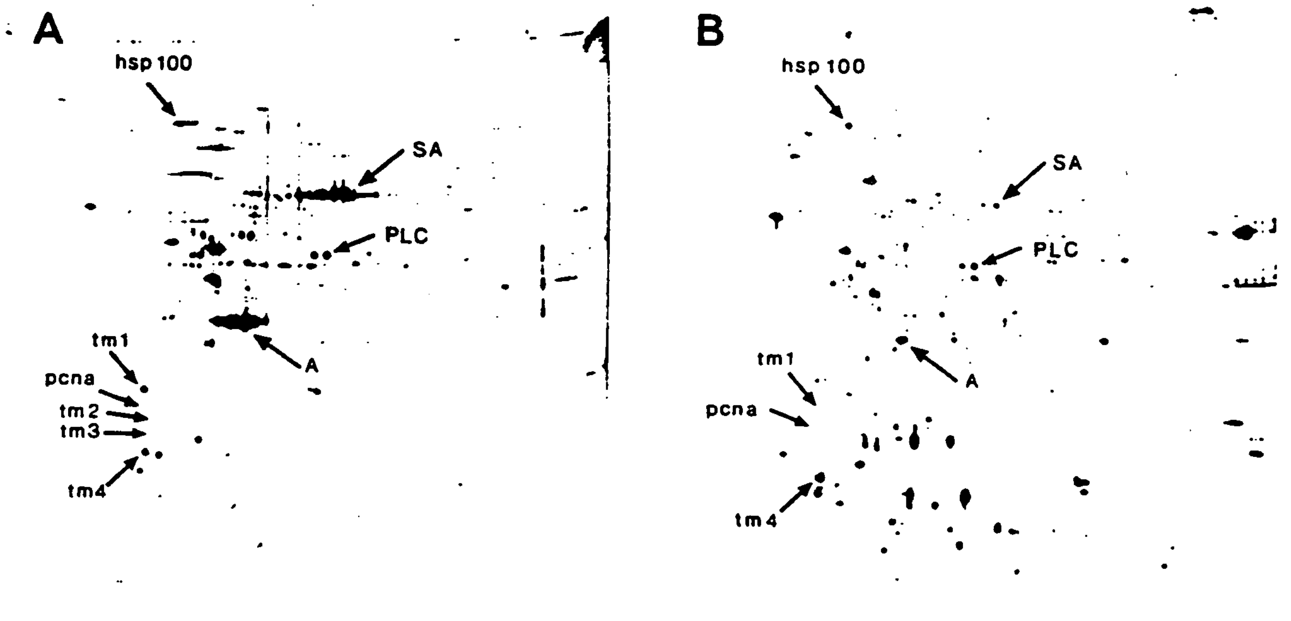


Figure 3. 2-DE analysis of a case of lung adenocarcinoma (LA). Comparison of 2-DE gel quality between (A) frozen and (B) fresh (needle aspiration) tissue preparation.

cells (Fig. 2d). Polypeptides were identified through a laboratory exchange of cell samples/2-DE maps and through 2-DE analysis of purified proteins (Table 1).

3.2 Preparation of samples from solid tumors

3.2.1 Fresh *versus* frozen tissue

An adenocarcinoma of the lung (LA) was prepared for 2-DE by conventional methods using frozen material (Fig. 3a). There are several possibilities for the poor resolution using frozen tissue, including the presence of high molecular weight protein aggregates. Filtering extracts through 0.1 µm filters (Durapore, Millipore) resulted in a slightly improved resolution (not shown). When fresh tumor tissue from tumor LA was used for sample preparation, using fine needle aspiration to collect the cells, the resolution was considerably improved (Fig. 3b). The use of fresh tissue resulted in a general increase in resolution, which was most pronounced in the 50–100 kDa molecular mass range. A number of differences in the protein profiles of the gels in Figs. 3a and 3b can be observed, some of which are indicated in the figures. The decrease in serum albumin in Fig. 3b is likely to result from loss of serum proteins occurring when cells were pelleted after aspiration. Other differences, such as the decreased level of transformation-sensitive tropomyosins (TM1-TM3), may result from enrichment of tumor cells in the sample of Fig. 3b. Fine needle aspiration, a well-established technique in cytology, extracts mainly tumor cells because of decreased intercellular adhesiveness of neoplastic cells as compared to normal tissue. Microscopic examination of Diff-Quick-stained extracted cells from case LA revealed almost 100% tumor cells, whereas the whole tissue extract contained approximately 60% tumor cells.

Table 1. Names and abbreviations for identified spots

Spot	Name	Basis for identification
A	Actins	a
aA	*alpha*-Actinin	a
B23	Protein B23 /Numatrin	a
EF2	Elongation factor 2	a
EF1	Elongation factor 1 β	a
GT	Glutathione-S-transpherase (*pi*	a
hsp60	Heat shock protein 60	a
hsp73	Heat shock protein 73	a
hsp80	Heat shock protein 80, GRP78, BIP	a
hsp90	Heat shock protein 90	a
hsp100	Heat shock protein 100, Endoplasmin	a
IFa	Intermediary filament associated	a
k8	Cytokeratin 8	b and a
LamB	Lamin B	a
Lip1	Lipocortin I	a
Lip2	Lipocortin II	a
Lip5	Lipocortin V	a
Mit1	Mitcon 1/β – F1 ATPase	a
Mit2	Mitcon 2	a
Mit3	Mitcon 3	a
MRP	Mucine Related Polypeptides	–
pcna	Ploliferating cell nuclear antigen	c and a
PLC	Phospholipase C (1)	a
RO	RO/SS-A antigen	a
SA	Serum Albumin	b and a
aT	*alpha*-Tubulin	a
bT	*betha*-Tubulin	a
tm1	Non-muscle tropomyosin isoform 1	b and a
tm2	Non-muscle tropomyosin isoferm 2	b and a
tm3	Non-muscle tropomyosin isoferm 3	b and a
tm4	Non-muscle tropomyosin isoform 4	b and a
tm5	Non-muscle tropomyosin isoform 5	b and a
TPI	Triose phosphate isomerase	a
V	Vimentin	b and a
Vid1	Vimentin derived protein	b and a
Vid2	Vimentin derived protein	b and a
Vid3	Vimentin derived protein	b and a
Vid4	Vimentin derived protein	b and a
Vin	Vinculin	a

a. homologous position with respect to other mammalian systems
b. purified protein(s)
c. immunoblotting

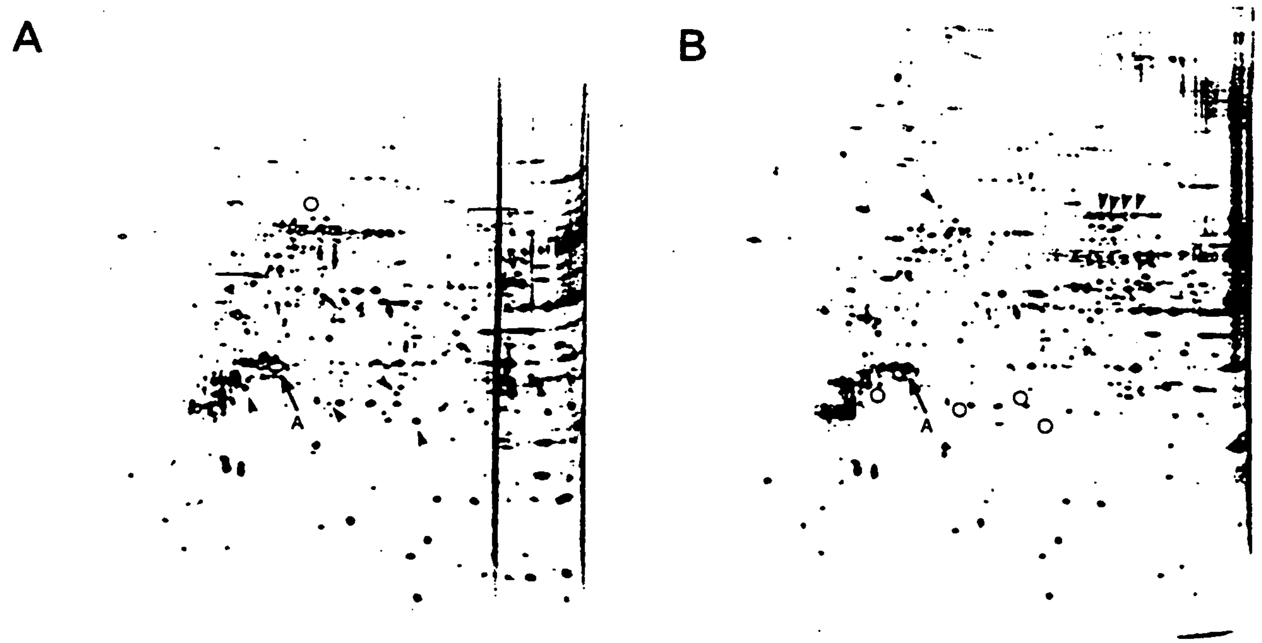


Figure 4. 2-DE analysis of a case of breast carcinoma (BC). Comparison of 2-DE quality and some differences in detected spots (arrow heads indicate increased intensity and circles or bracket indicate decreased intensity of the same spots) between (A) enzymatically and (B) nonenzymatically (scraped) tissue preparation.

3.2.2 Comparison of different methods for preparing cells from fresh tumor tissue

Samples were prepared from breast and lung carcinomas using either an enzymatic treatment with collagenase/elastase or using nonenzymatic preparations (Fig. 4). A number of differences in the protein profiles were observed in the resulting 2-DE gels, some of which are indicated in Figs. 4a and b. These differences include both increases and decreases in spot intensity. These differences may result from degradation of high molecular weight polypeptides during enzymatic treatment, increased solubilization of polypeptides, or may have other causes. For many tumors, it was only possible to obtain small amounts of material since they were reserved for other examinations. In these cases, samples could be prepared for 2-DE using either needle aspiration or scraping. Figure 5a shows a 2-DE gel prepared from squamous lung carcinoma (LS) cells collected by needle aspiration and Fig. 5b shows a gel prepared from the same tumor by scraping. In this case, a number of differences were recorded between the two procedures, some of which are arrowed in Fig. 5. Samples obtained from other tumors (breast and lung) generally showed fewer differences between these two methods of cell sampling (not shown). These data show that different nonenzymatic extraction procedures may yield different polypeptide patterns. However, the number of spots with a large

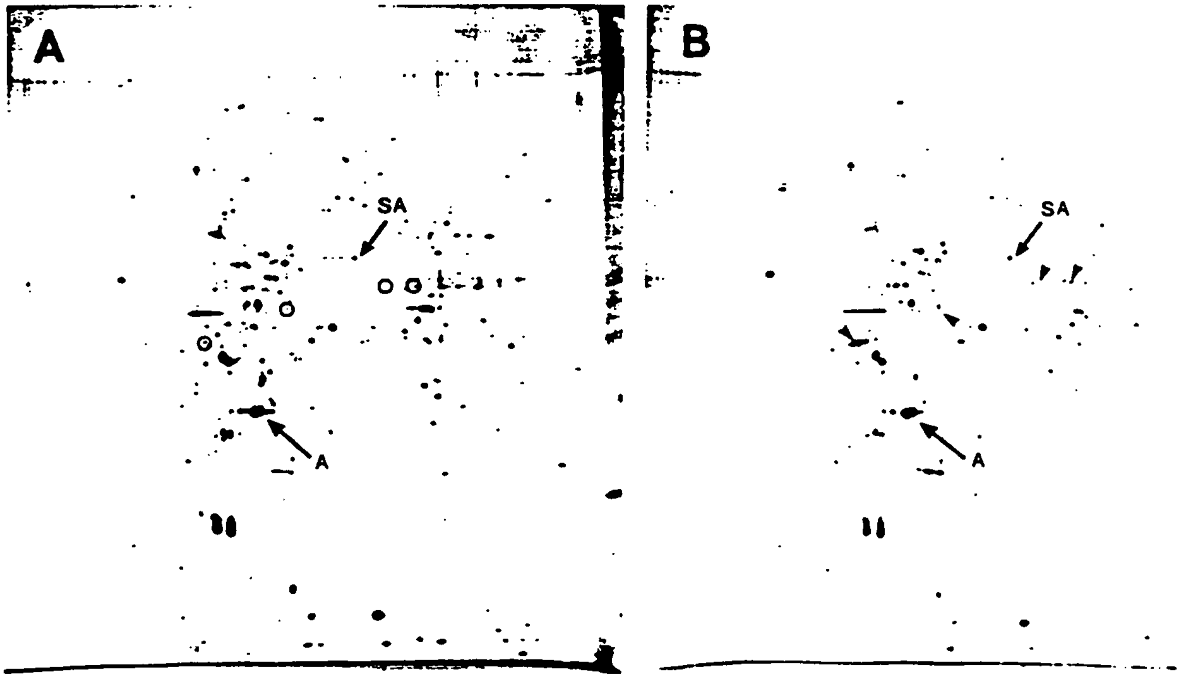


Figure 5. 2-DE analysis of a case of lung cancer (LS). Comparison of 2-DE gel quality and detected spots (arrow heads and circles) between (A) aspirated (needle aspiration) and (B) scraped preparations from fresh tissue.

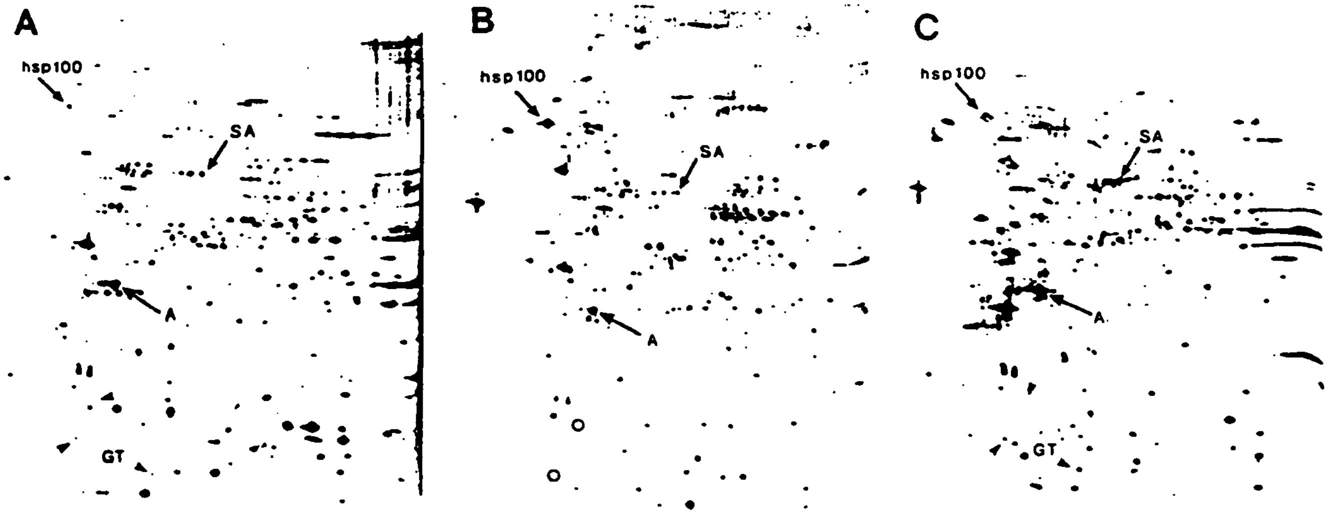


Figure 6. 2-DE analysis of three other types of tumors, (A) hypernephroma, (B) an adenoma of the thyroid and (C) corpus cancer, using the nonenzymatic preparation technique. Arrowheads and circles indicate some cytosolic polypeptides.

difference in intensity were lower than when a nonenzymatic preparation was compared with an enzymatic preparation.

2-DE maps of satisfactory quality were prepared by a third procedure. Cells were released from small pieces of tumor by squeezing (see Section 2). Some examples of this are shown in Fig. 6 where 2-DE maps derived from a case of hypernephroma, KH (Fig. 6a), a case of thyroid tumor, TA (Fig. 6b) and a case of corpus cancer, CP (Fig. 6c) can be seen. We conclude that nonenzymatic techniques are useful for 2-DE analysis of a number of different tumors. The quality of the resulting gels is comparable to that obtained using cultured cells (compare the gels in Fig. 2 with those in Fig. 4, 6 and 7). Which of these methods will be optimal will, in our experience, depend on the tumor material. For example, very small tumors are preferably extracted by squeezing; on the other hand, breast cancers (which are often fibrous) yield satisfactory samples using scraping.

3.2.3 Purification of cells on percoll gradients

We considered the possible advantage of separating viable cells from dead cells, erythrocytes, and debris using discontinuous Percoll gradients. Cells collected

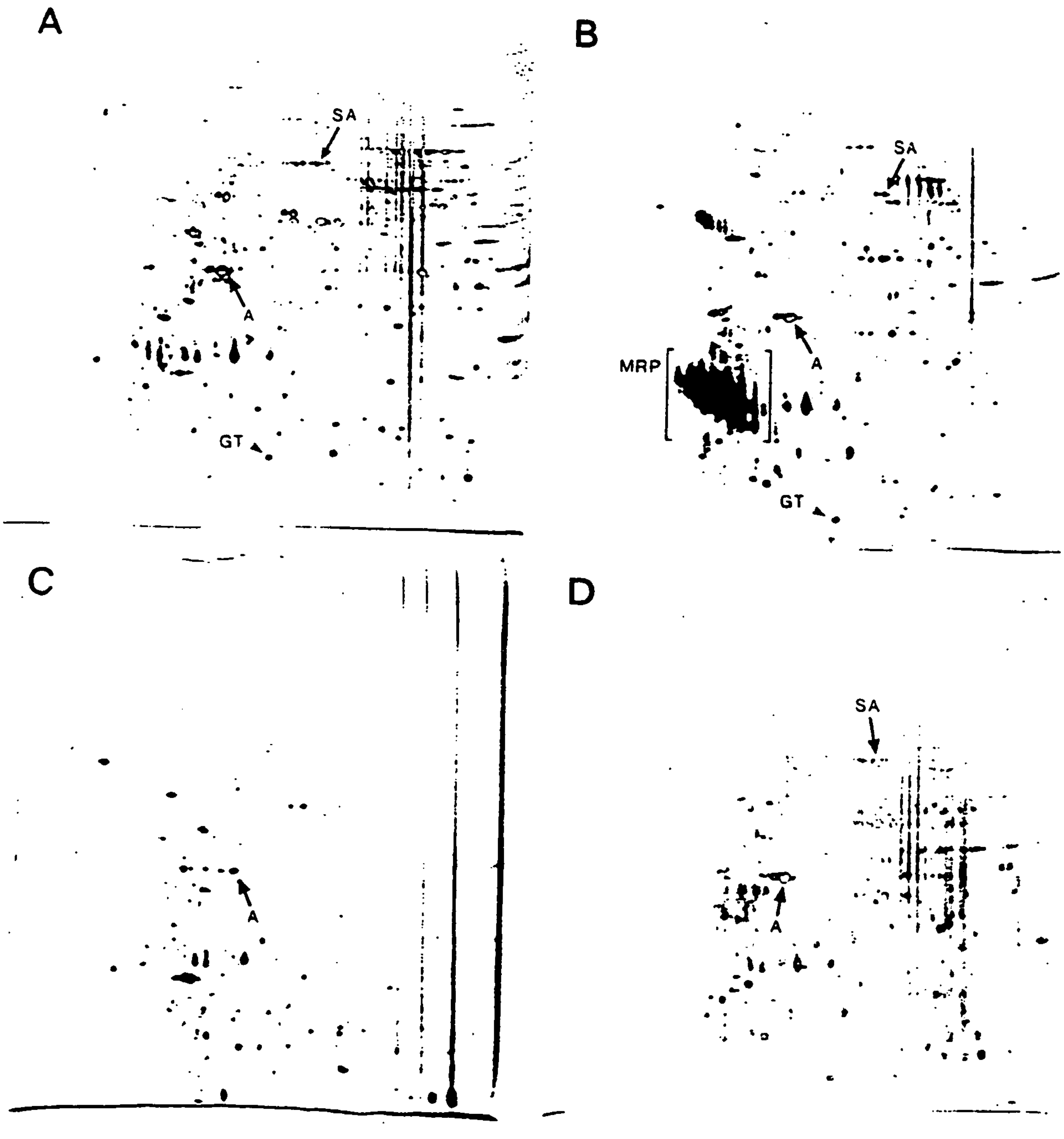


Figure 7. 2-DE analysis of polypeptides from viable (b and d) and nonviable (a and c) cells of an adenocarcinoma of the lung (LB), separated using discontinuous Percoll density gradient. Nonenzymatic preparation technique (a and b) and enzymatic preparation technique (c and d) are compared.

from the interphase showed a viability of more than 90% as judged by trypan blue exclusion test. However, it as found that the yield of viable cells decreased dramatically if the tissue resection was not immediately processed. To study the effect of lysis of cells during the preparation procedure, 2-DE maps were prepared from nonenzymatically extracted cells of case LB collected from the top fraction (nonviable, Fig. 7a) and interphase fraction (viable, Fig. 7b). These 2-DE maps were compared with corresponding fractions (nonviable, Fig. 7c, and viable, Fig. 7d) of enzymatically extracted cells. One clear disadvantage of the enzymatic technique was that when loss of cell viability occurred during preparation, a dramatic loss of high molecular weight polypeptides was observed (Fig. 7c). This was probably due to degradation of intracellular proteins. However, nonenzymatic preparations showed fewer differences between viable and nonviable cells: The most pronounced alteration was a decrease of a group of mucine related proteins (Fig. 7b). We conclude, therefore, that discontinuous Percoll gradient is necessary after enzymatic extraction of cells, but can be omitted from the nonenzymatical tumor sample preparation procedure.

We used the MDA-231 cell line to study the effects of cell lysis and leakage of cytosolic polypeptides during sample preparation. Remarkably, after 30, 50, 80 and 140 min of incubation in PBS/PIH at 0°C, no significant changes were observed in the 2-DE pattern (not shown). Although loss of cell viability may not result in protein degradation when cells are incubated in the presence of protease inhibitors, loss of cytosolic proteins would be expected during pelleting of cells. We monitored the loss of lactate dehydrogenase (LDH) activity into the supernatant during incubation in PBS of MDA-231 and MCF-7 breast cancer cells at 20°C. In both cases, loss of viability was paralleled by release of LDH from the cells (Fig. 8). After 5 h, 70% of the MCF-7 cells, but only 30% of the MDA-231 cells were dead (not shown).

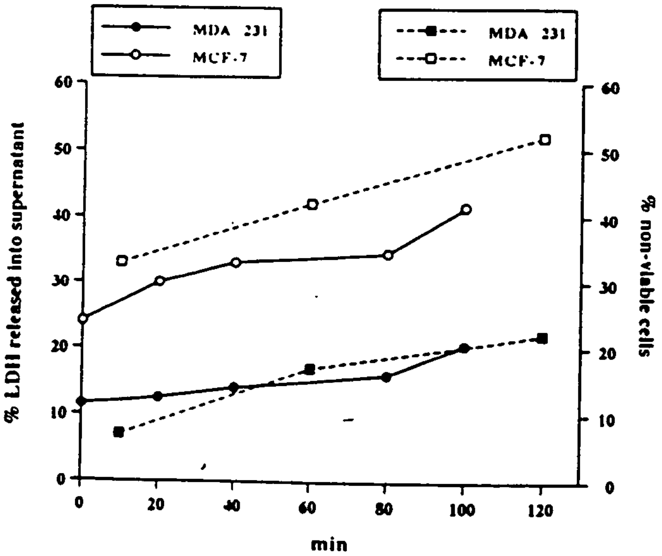


Figure 8. The relative release (fraction in supernatant of total) of lactate dehydrogenase activity (LDH) and cella viability *versus* incubation time of the mammary carcinoma cell lines MDA-231 and MCF-7 during incubation in PBS at 20°C.

These data indicate the impact of a rapid preparation procedure, at low temperature, of fresh tumor samples. Experiments have also been performed using only 1.07 g/mL Percoll (Fig. 6c and Fig. 1, left test tube) in order to remove erythrocytes. One clear advantage with this procedure, which today is routinely utilized, is a higher yield of viable cells, probably due to decreased sample preparation time.

4 Discussion

We describe procedures for sample preparation from solid tumors for 2-DE. 2-DE maps could be derived from solid tumors which were similar in quality to those obtained from cultured cells. Compared to methods using frozen material, the resolving power of the 2-DE technique is increased, allowing examination of a large number of polypeptides from tumors of different malignancies. Other investigators [12,22] have used samples from frozen tumors to derive 2-DE maps. We have previously described disadvantages encountered using frozen tumor samples including variations in contaminating proteins between different samples [3]. The methods described here are based on the preparation of cells from tumors without enzymatic digestion. The enzymatic step could be avoided since malignant cells usually grow as solid masses which are not strongly attached to the matrix. Furthermore, we found that omitting the enzymatic digestion alleviated the necessity of purifying viable tumor cells on Percoll gradients. This was in sharp contrast to enzymatically treated samples, where loss of viability leads to loss of high molecular weight proteins (Fig. 7c).

At least in the case of lung cancer, viable and nonviable cells showed small differences in respect to 2-DE maps. Presumably, protease inhibitors penetrate cells and inhibit proteolysis. In model experiments, we observed leakage of cytosolic protein (LDH) from the cells in parallel to loss of viability. Apparently, however, only a limited decrease of the level of low molecular weight cytosolic polypeptides was detected using silver staining combined with visual inspection. We have found that although some tumors are well suited for the preparation procedure described, others are not. In general, good results were obtained using tumors of the lung, breast, corpus and lymphomas. In contrast, cells from thyroid adenomas and hypernephroma showed poor viability. We were in these cases unable to separate nonviable cells from viable cells, and we can therefore not evaluate the consequence of the loss of viability on 2-DE patterns, apart from a loss of some low molecular weight cytosolic polypeptides.

Highly differentiated tumors may show lower viability as compared with poorly differentiated tumors (Dr. Farkas Vanky, personal communication). A number of samples from thyroid tumors were prepared for 2-DE but most cases showed poor viability. We believe that special care is needed during preparation of generally highly differentiated tumor groups. The difference between loss of viability/leakage of LDH of the more differentiated MCF-7 cells and the less differentiated MDA-231 cells is in line with these observations (Fig. 8). A number of potential and interesting markers, like tropomyosin isoforms, cytokeratins and heat shock proteins, appear to be insensitive to loss of viability during the preparation procedure. We have to date made numerous observations of alterations in the expression of these polypeptides in breast cancers and lung cancers.

Another problem that may occur, irrespective of sample preparation techniques used, is admixture of lymphocytes. These cases are easily detectable in smears and it may therefore be possible to select lymphocyte specific spots as "internal markers" for the 2-D PAGE analysis. Studies using this approach are in progress. Many of the polypeptides identified are structural (Table 1). Since the expression of many of these polypeptides are known to vary between normal and malignant cells, the possibility to determine their expression simultaneously is appealing. In the specific case of breast cancer, alterations in the expression of intermediate filament proteins (cytokeratins) are known to occur during tumor progression [23]. Other proteins known to be differentially expressed between normal cells and transformed cells are tropomyosins, numatrin/B23, heat shock proteins and PCNA. To this end, we have observed alterations in the expression of cytokeratin 8, hsp 90, and non-muscle tropomyosin isoform 2 during malignant progression. (Okuzawa *et al.*, in preparation and Franzén *et al.*, in preparation).

The method of choice for sample preparation from tumor tissues will depend on the properties of the tumor material studied. It may be important to use only one method when comparing cases within one group, as differences were observed between methods. The advantages of the nonenzymatic techniques are (i) that it minimizes contamination with connective tissue, (ii) that problems with contamination of serum proteins are avoided, and (iii) that separation of viable and dead cells is not necessary. Hereby the revolving power of 2-D PAGE is maximized for the analysis of human tumors and studies on inter-tumor variations in gene expression are facilitated. In addition, the polypeptide patterns obtained may be more representative for the *in vivo* tumor cell since the use of enzymes and incubations have been minimized.

We would like to thank Dr. J. I. Garrels, Dr. S. Pattersson, Dr. S. M. Hanash and Dr. J. E. Celis for making sample and 2-DE map exchanges possible. This study was supported by grants from the Swedish Cancer Society and the Cancer Society in Stockholm.

Received March 5, 1993

5 References

[1] Celis, J. E., Dejgaard, K., Madsen, P., Leffers, H., Gesser, B., Honore, B., Rasmussen, H. H., Olsen, E., Lauridsen, J. B. and Ratz, G., *Electrophoresis* 1990, *11*, 1072–1113.

[2] Garrels, J. I., Franza, B. R., Chang, C., Latter, G., *Electrophoresis* 1990, *11*, 1114–1130.

[3] Franzén, B., Iwabuchi, H., Kato, H., Lindholm, J. and Auer G., *Electrophoresis 1991, 12*, 509–515.

[4] Sherwood, E. R., Berg, L. A., Mitchell, N. J., McNeal, J. E., Kozlowski, J. M. and Lee, C., *J. Urology* 1990, *143*, 167–171.

[5] Endler, A. T., Young, D. S., Wold, L. E., Lieber, M. M. and Curie, R. M., *J. Clin. Chem. Clin. Biochem. 1986, 24*, 981–992.

[6] Forchhammer, J. and Macdonald-Bravo, H., in: Celis, J. E. and Bravo, R., (eds.), *Gene Expression in Normal and Transformed Cells*, Plenum, New York 1983, pp. 291–314.

[7] Linder, S., Brzeski, H. and Ringertz, N. R. *Exp. Cell. Res.* 1979, *120*, 1–14.

[8] Celis, J. E. and Bravo, R. (Eds.), *Two-dimensional Gel Electrophoresis of Proteins*, Academic Press, New York 1984, pp. 3–36.

[9] Garrels, J. I., *J. Biol. Chem.* 1979, *254*, 7961–7977.

[10] Anderson, N. L., *Two-Dimensional Electrophoresis. Operation of the ISO-DALT System*. Large Scale Biology Press, Washington, DC 1988, 162.

[11] Bradford, M., *Anal. Biochem.* 1976, *72*, 248.

[12] Tracy, R. P., Wold, L. E., Currie, L. M. and Young, D. S., *Clin. Chem.* 1982, *28*, 890–899.

[13] Merril, C. R., Goldman, D., Sedman, S. A. and Elbert, H. M., *Science* 1981, *211*, 1437–1438.

[14] Morrissey, J. H., *Anal Biochem.* 1981, *117*, 307–310.

[15] Gard, D. L., Bell, P. B., Lazarides, E., *Proc. Natl. Acad. Sci. USA*, 1979, *76*, 3894–3898.

[16] Matsumura, F., Lin, J.-C., Yamashiko-Matsumura, S., Thomas, G. P. and Topp, W. C., *J. Biol. Chem.*, 1983, *258*, 13954–13960.

[17] Paulin, D., Forest, N. and Perreau, J., *J. Mol. Biol.* 1980, *144*, 95–101

[18] Blobel, G. A., Moll, R., Franke, W. W., Kayser, K. W. and Gould, V. E., *Am. J. Pathol.* 1985, *121*, 235–247.

[19] Ochs, D. C., McConkey, H. E. and Guard, N. L., *Exp. Cell. Res.* 1981, *135*, 355–362.

[20] Bhattacharya, B., Gaddamanuga, L.P., Valverius, E. M., Salomon, D. S. and H. L. Cooper, *Cancer Res.* 1990, *50*, 2105–2112.

[21] Sommers, C. L., Walker-Jones, D., Heckford, S. E., Worland, P., Valverius, A., Clark, R., McCornick, F., Stampfer, M., Abularch, S. and Gelmann, E. P., *Cancer Res.* 1989, *49*, 4258–4263.

[22] Trask, D. K., Band, V., Zajchwski, D. A., Yaswen, P., Suh, T. and Sager, R., *Proc. Natl. Acad. Sci. USA* 1990, *87*, 2319–2323.

[23] Trask, D. K., Bond, V., Zajchanski, D. A., Yaswen, P., Suh, T. and Sager, R., *Proc. Natl. Acad. Sci. USA* 1990, *87*, 2319–2323.

F

Electrophoresis 1994, 15, 529–539

Bengt Bjellqvist*
Bodil Basse
Eydfinnur Olsen
Julio E. Celis

Institute of Medical Biochemistry and Danish Centre for Human Genome Research, Aarhus University, Aarhus

Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions

A highly reproducible, commercial and nonlinear, wide-range immobilized pH gradient (IPG) was used to generate two-dimensional (2-D) gel maps of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human epidermal keratinocytes. Forty one proteins, common to most human cell types and recorded in the human keratinocyte 2-D gel protein database were identified in the 2-D gel maps and their isoelectric points (p*I*) were determined using narrow-range IPGs. The latter established a pH scale that allowed comparisons between 2-D gel maps generated either with other IPGs in the first dimension or with different human protein samples. Of the 41 proteins identified, a subset of 18 was defined as suitable to evaluate the correlation between calculated and experimental p*I* values for polypeptides with known composition. The variance calculated for the discrepancies between calculated and experimental p*I* values for these proteins was 0.001 pH units. Comparison of the values by the *t*-test for dependent samples (paired test) gave a *p*-level of 0.49, indicating that there is no significant difference between the calculated and experimental p*I* values. The precision of the calculated values depended on the buffer capacity of the proteins, and on average, it improved with increased buffer capacity. As shown here, the widely available information on protein sequences cannot, *a priori*, be assumed to be sufficient for calculating p*I* values because post-translational modifications, in particular *N*-terminal blockage, pose a major problem. Of the 36 proteins analyzed in this study, 18–20 were found to be *N*-terminally blocked and of these only 6 were indicated as such in databases. The probability of *N*-terminal blockage depended on the nature of the *N*-terminal group. Twenty six of the proteins had either M, S or A as *N*-terminal amino acids and of these 17–19 were blocked. Only 1 in 10 proteins containing other *N*-terminal groups were blocked.

1 Introduction

As compared with carrier ampholyte isoelectric focusing (CA-IEF), the application of immobilized pH gradients (IPGs) in the first dimension in 2-D gel electrophoresis offers improved reproducibility [1] because the nature of the pH gradient makes the resulting focusing positions insensitive to the focusing time [2] and to the type of sample applied [3]. The recently introduced ready-made IPG strips [4] seem to be an ideal substitute for the carrier ampholyte gradients, which until now have been the most commonly used first dimensions in 2-D gel electrophoresis. The availability of standardized first dimensions opens the possibility of comparing 2-D gel maps of various cell types generated in different laboratories, provided that the focusing positions of a number of easily recognizable polypeptide spots common to the cell types in question are known. Even though this approach is limited to experiments performed with the same standardized IPG, the flexibility provided by IPGs allows the pH gradient to be adjusted to the requirements of a particular experiment.

Exchange and communication of 2-D gel protein data requires a pH scale that is independent of the particular IPG used and by which the results can be described. The introduction of carbamylation trains and the relation of focusing positions to the spots in these trains represented a step forward towards solving the reproducibility problem experienced with carrier ampholyte focusing [5]. Problems associated with the use of carbamylation trains were mainly due to lack of temperature control and to the use of nonequilibrium focusing conditions. Accordingly, the pattern variation involved not only the resulting pH gradients, but also the relative spot positions as related to each other and to spots in the carbamylation trains. Even though the question of reproducibility has, to a large extent, been solved, the carbamylation trains are still not ideal as markers because the spots in the trains do not represent defined entities but rather a large number of differently carbamylated peptides having close p*I* values. As a result, the spots are large and poorly defined as compared to the ordinary polypeptide spots in 2-D gel maps.

Correspondence: Professor J. E. Celis, Institute of Medical Biochemistry and Danish Centre for Human Genome Research, Aarhus University, DK-8000 Aarhus C, Denmark

Abbreviations: CA-IEF, carrier ampholyte-isoelectric focusing; SSP, sample spot number

* Present addess: Pharmacia Biotech AB, S-751 82 Uppsala, Sweden

0173-0835/94/0304-0529 $5.00+.25/0

Neidhardt *et al.* [6] defined the pH gradient in 2-D gel experiments by p*I* markers whose p*I* values were calculated from the amino acid composition. Focusing positions of other polypeptides could be predicted from their composition but the p*K* values needed for the p*I* calculations were unknown. Various groups employing this approach do not use the same pK values [6, 7] and therefore, the p*I* values derived in this way cannot be expected to describe the variation of the hydrogen ion activity. In spite of this fact, it is still possible to make approximate predictions of focusing positions because the p*K* values used to define the pH gradient are also used to calculate p*I* values and to predict the focusing positions. Errors in p*K* assignments are therefore compensated. A pH scale which corretly reflects the variation in hydrogen ion activity during focusing should improve the precision of the predictions, but this has never been implemented with CA-IEF focusing as a first dimension in 2-D gel electrophoresis. The main reason for this are the problems associated with pH measurements in focused gels containing high concentrations of urea.

IPGs can be described from the concentration variation of the immobilized groups, provided that the p*K* values of these groups are known for the conditions prevailing during focusing. To avoid measurements on gels, Gianazza *et al.* [8] suggested the use of p*K* values derived by addition of determined p*K* shifts. Recently, direct determinations of p*K* differences between immobilized groups in IPGs were made by determining p*I*-p*K* values in overlapping narrow-range IPGs [9, 10] and the results verified the applicability of the Gianazza approach. A description of the focusing results in a pH scale, which correctly describes the variation of the hydrogen ion activity for the focusing conditions used, not only allows the comparison of 2-D gel maps generated with different IPGs, but also opens the possibility for correlating the focusing position of a polypeptide with its composition [9]. Experiments by Bjellqvist *et al.* [9, 10] have implied that pH scales showing good correlation between calculated and experimental p*I* values can be derived for any of the conditions commonly used for focusing in connection with 2-D gel electrophoresis. These pH scales are then defined through the p*K* values of the immobilized groups in the IPG containing gel. To be useful for interlaboratory comparisons, however, the pH scale has to be defined through p*I* values of easily recognizable spots present in the 2-D gel map. So far, p*I* determinations in a useful pH scale, combined with determinations of pK values needed for p*I* calculations, have only been made for the pH range 4.5–6.5 at 10°C [9]. CA-IEF focusing as described by O'Farrell [11] does not control the temperature of the first dimension, which can be expected to be slightly above room temperature. With IPGs, the temperature commonly used is about 20°C [4, 12] or 25°C [13] and this is a critical parameter that needs to be controlled [14].

The present work was designed to compare 2-D gel maps of different cell types in a laboratory applying both CA-IEF and IPG focusing at a common temperature. To this end we have generated 2-D gel maps of proteins from noncultured, unfractionated normal human epidermal keratinocytes with IPG in the first dimension and a focusing temperature of 25°C. We have used commercial nonlinear, wide-range IPG strips which give 2-D gel maps that are closely similar to the ones resulting with the CA-IEF technique used to establish the human keratinocyte database [15]. As an initial step towards interlaboratory comparisons of results obtained with the nonlinear gradient as a first dimension we report here on the focusing positions of 41 known proteins that are common to most human cell types. The pH range covered corresponds to the range in classical CA-IEF 2-D gel electrophoresis and in order to use these proteins as internal standards for comparing 2-D gel maps generated with other IPGs we determined their p*I* values with narrow-range IPGs in the first dimension. We have compared the calculated *versus* experimental p*I* values and show that it is necessary to have further information (absence or presence and nature of posttranslational modifications), in addition to amino acid composition to be able to calculate p*I* values that correspond to the actual experimental values. The p*K* values used for the calculations are provided and the usefulness of p*I* prediction in relation to database information is discussed. Furthermore, we comment on the possibility of using experimentally determined p*I* values to verify the available database information on polypeptide composition.

2 Materials and methods

2.1 Apparatus and chemicals

Equipment for isoelectric focusing and horizontal SDS electrophoresis (Multiphor® II electrophoresis chamber, Immobiline® strip tray, Multidrive XL programmable power supply, Macrodrive power supply and Multitemp® II) was from Pharmacia LKB Biotechnology AB (Uppsala, Sweden). Vertical second-dimensional gels were run in the home-made equipment described in [15]. The IPG strips with the wide-range nonlinear pH gradient were either Immobiline DryStrip® pH 3–10 NL, 180 mm or alternatively 160 mm long IPG strips with a corresponding pH gradient. In both cases the IPG strips were delivered by Pharmacia LKB. Immobiline, Pharmalyte, Ampholine, GelBond as well as PAG film and the ready-made horizontal SDS gels (ExcelGel® XL SDS 12–14) were also from Pharmacia LKB. Purified proteins and peptides were from Sigma (St. Louis, MO).

2.2 Sample preparation

Preparation and labeling of unfractionated keratinocytes as well as fibroblasts have been described in [16]. Cells were lysed in a solution containing 9.8 M urea, 2% w/v NP-40, 100 mM DTT and 2% v/v Ampholine pH 7–9.

2.3 2-D gel electrophoresis

First-dimensional focusing was performed according to Görg *et al.* [2] with some minor modifications, as described in [9]. Rehydration of the IPG strips was made in a solution containing 9.8 M urea, 2% w/v CHAPS, 10 mM DTT and 2% v/v carrier ampholyte mixture. The carrier ampholyte mixture consisted of 2 parts Pharmalyte

4–6.5, 1 part Ampholine pH 6–8 and 1 part Pharmalyte pH 8–10.5. Usually, cathodic sample application was used and the samples were diluted 2–20 times in a solution containing 9.8 M urea, 4% w/v CHAPS, 1% w/v DTT and 35 mM Tris base. For acidic application, the Tris-base was substituted with 100 mM acetic acid. The degree of dilution and sample volume (20–100 µL) depended on the particular sample and the IPG, and whether visualization of the proteins was to be done by Coomassie Brilliant Blue or silver staining. With the wide-range non-linear IPG, 10–30 µg of total protein was loaded for silver staining and 100–200 µg for Coomassie staining. Focusing was done overnight with Vh products in the range of 45–60 kVh with 160 mm long strips and 50–70 kVh with 180 mm long strips. Solubilization of polypeptides and blocking of -SH groups prior to the second-dimensional run, as well as loading on the second-dimensional gel was done as described in [9]. The stacking gel was omitted and 5–10 mm were left at the top of the second-dimensional gel for applying the IPG strip. The space was filled with electrode buffer containing 0.5% w/v agarose. Casting, running, staining and autoradiography were carried out as described in [15].

2.4 Experimental determination of p*I* values

The determination of the p*K* differences between Immobilines p*K* 4.6, p*K* 6.2 and p*K* 7.0 necessary for the calibration of the pH scale at 25°C in 9.8 M urea was done as described in [9] with the same narrow-range IPGs. The pH scale was defined by setting the p*K* value of Immobiline p*K* 4.6 equal to 4.61 [9] and the determined p*K* differences gave the p*K* values of Immobilines p*K* 6.2 and p*K* 7.0, equal to 5.73 and 6.54, respectively. The p*K* differences found are in good agreement with values derived from [17] and [8] by extrapolation to 9.8 M urea concentration. As in [9], additional narrow-range recipes have been used for determining p*I* values. With narrow-range IPGs extending to pH values higher than the p*K* value of Immobiline p*K* 7.0, anodic sample application was used with acetic acid added to the sample solution. Otherwise, cathodic sample application was used with the same sample buffer as for wide-range IPGs.

2.5 Protein compositions used for p*I* calculations

With the exception of vimentin, protein compositions are from the Swiss-Prot database [18]. For vimentin, we used the data from [19], where the amino acid at position 41 is a D instead of a S. Information in the Swiss-Prot database on phosphorylation has been disregarded because it was known from earlier studies (J. E. Celis, unpublished results) that the spots in question corresponded to the unphosphorylated forms of the peptides.

2.6 Calculation of p*I* values

For the p*I* calculations it was assumed that the same p*K* value could be used for an amino acid residue in all polypeptides and in all positions in the peptide except for *N*- or *C*-terminally placed amino acids. For the p*K* values of the *N*-terminal amino groups the effect of the different substituents on the α-carbon were taken into account. The calculations of p*I* values were made with the aid of the IPG-maker program [20].

2.7 p*K* values used for p*I* calculations

For the carboxyl terminal group and internal glutamyl and aspartyl residues the same p*K* values were used as in [9]. For *C*-terminal glutamyl and aspartyl residues, separate p*K* values were derived with the aid of the Taft equations [9, 21]. The p*K* values of histidyl groups were calculated from the p*I* values of human carbonic anhydrase I as in [9]. For *N*-terminal glycine a p*K* value of 7.50 was used. The p*K* shift caused by a substituent on the α-carbon was assumed to be identical with the p*K* shift the substituent caused for the amino group in the amino acid, *i.e.* 2.28 pH units were subtracted from the p*K* values for the amino groups in the amino acids given in [22, 23]. The approximate p*K* value of 9 for the cystenyl group was taken from [24]. For tyrosyl and arginyl groups we used the p*K* values for the amino acids [22, 23]. For lysyl groups the effect of high urea concentration on amino groups was taken into account and 0.5 pH units were subtracted from the amino acid p*K* value. These last three p*K* values are far from the pH range under study and the results found would have been the same if lysyl and arginyl groups were assumed to be fully ionized while the ionization of tyrosyl groups were neglected. A complete list of the p*K* values used is given in Table 1.

Table 1. p*K* Values used for the ionizable groups in peptides 9.8 M urea, 25°C

Ionizable group	p*K*
C-terminal	3.55*
N-terminal	
Ala	7.50
Met	7.00
Ser	6.93
Pro	8.36
Thr	6.82
Val	7.44
Glu	7.70
Internal	
Asp	4.05
Glu	4.45
His	5.98
Cys	9
Tyr	10
Lys	10
Arg	12
C-terminal side chain groups	
Asp	4.55
Glu	4.75

2.8 Statistical analysis

Statistical comparisons of the experimental and calculated p*I* values were done on an Apple Macintosh IIsi using the statistical package Statistica/Mac, release 3.0b (from StatSoft Inc., Tulsa, Oklahoma). Calculated and experimental p*I* values were compared by the *t*-test for correlated samples (paired *t*-test). The normality of p*I* differences was estimated graphically by probability plots. The variances of the data presented here and the similar data on plasma and liver proteins in [9] were compared by the F-test.

3 Results and discussion

3.1 Identification of polypeptides and p*I* determinations

The 2-D gel maps of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human keratinocytes, focused with the nonlinear, wide-range IPG and CA-IEF pH gradients in the first dimension, are shown in Figs. 1 and 2, respectively. The IPG extends to higher pH values but otherwise the two patterns are very similar and most of the spots in the IPG pattern can be directly related to the corresponding spots in the CA-IEF gel. To obtain comparable patterns it was important to keep the focusing temperature as similar as possible. Compared to other studies [1–4, 9, 10, 12–14], we increased the urea concentration in the focusing gel to 9.8 M because keratins streaked badly in the focusing dimension when 8 M urea was used, presumably due to

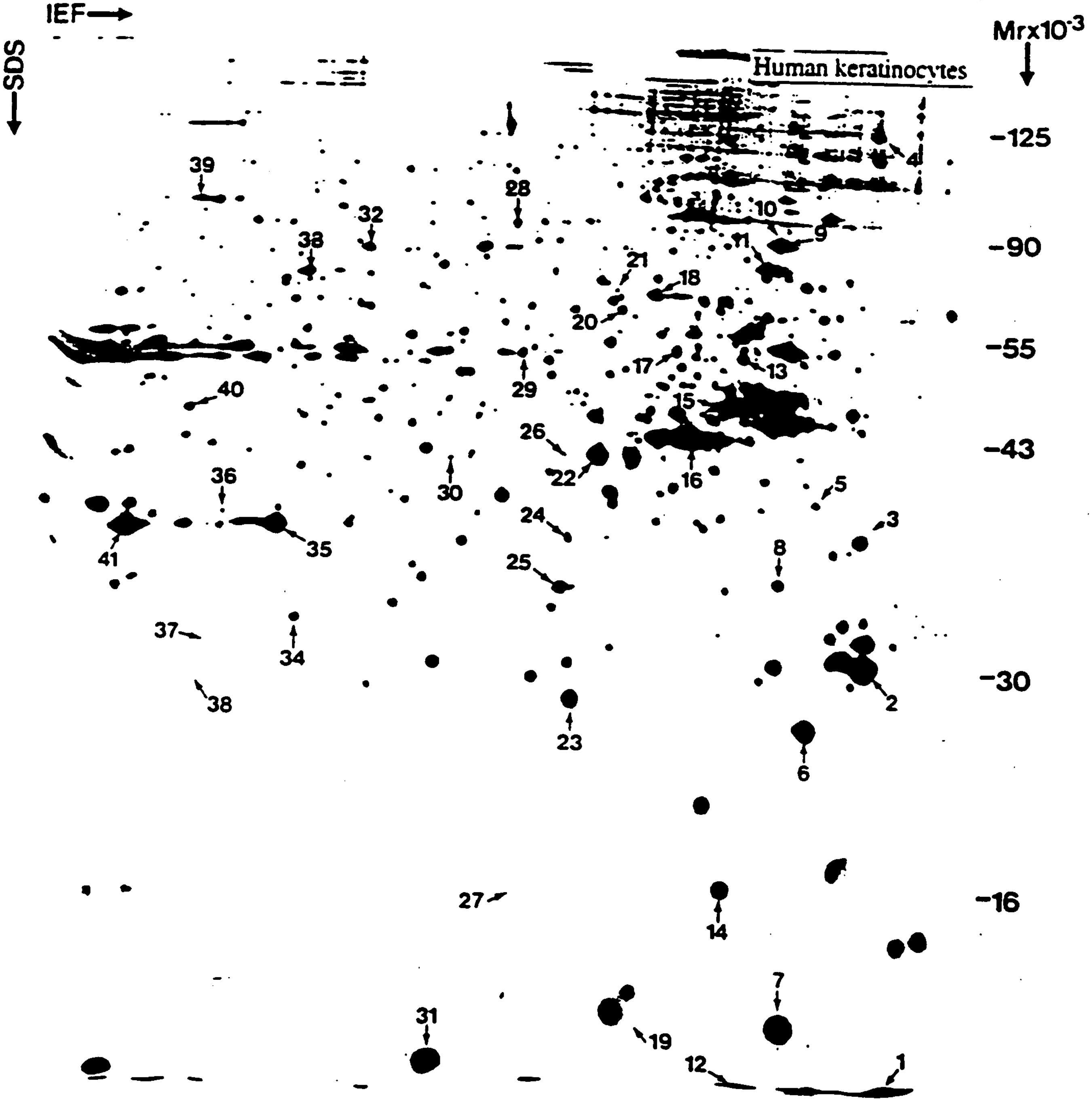


Figure 1. 2-D gel protein map of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human keratinocytes focused with the nonlinear, wide-range IPG in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

aggregates of acidic and basic keratins. An increase in urea concentration to 9 M or more eliminated these streaks: apart from this effect, no other major changes in the focusing positions were observed. In Fig. 1 we have indicated the positions of 41 known proteins from the human keratinocyte 2-D gel database that are most likely common to most human cell types. The choice was made because these proteins are easy to identify with certainty. With the exception of stratifin (spot 2), involucrin (spot 4) and keratin 14 (spot 15), which are all epithelial markers, these proteins are also present in human fibroblasts (Fig. 3) and lymphocytes (results not shown), and therefore can be used as landmarks for comparing 2-D gel maps derived from different cell types. In Table 2 the 41 proteins are listed together with their sample spot numbers (SSP) in the human keratinocyte protein database and p*I* values determined in 2-D gel maps generated with narrow-range IPGs in the first dimension.

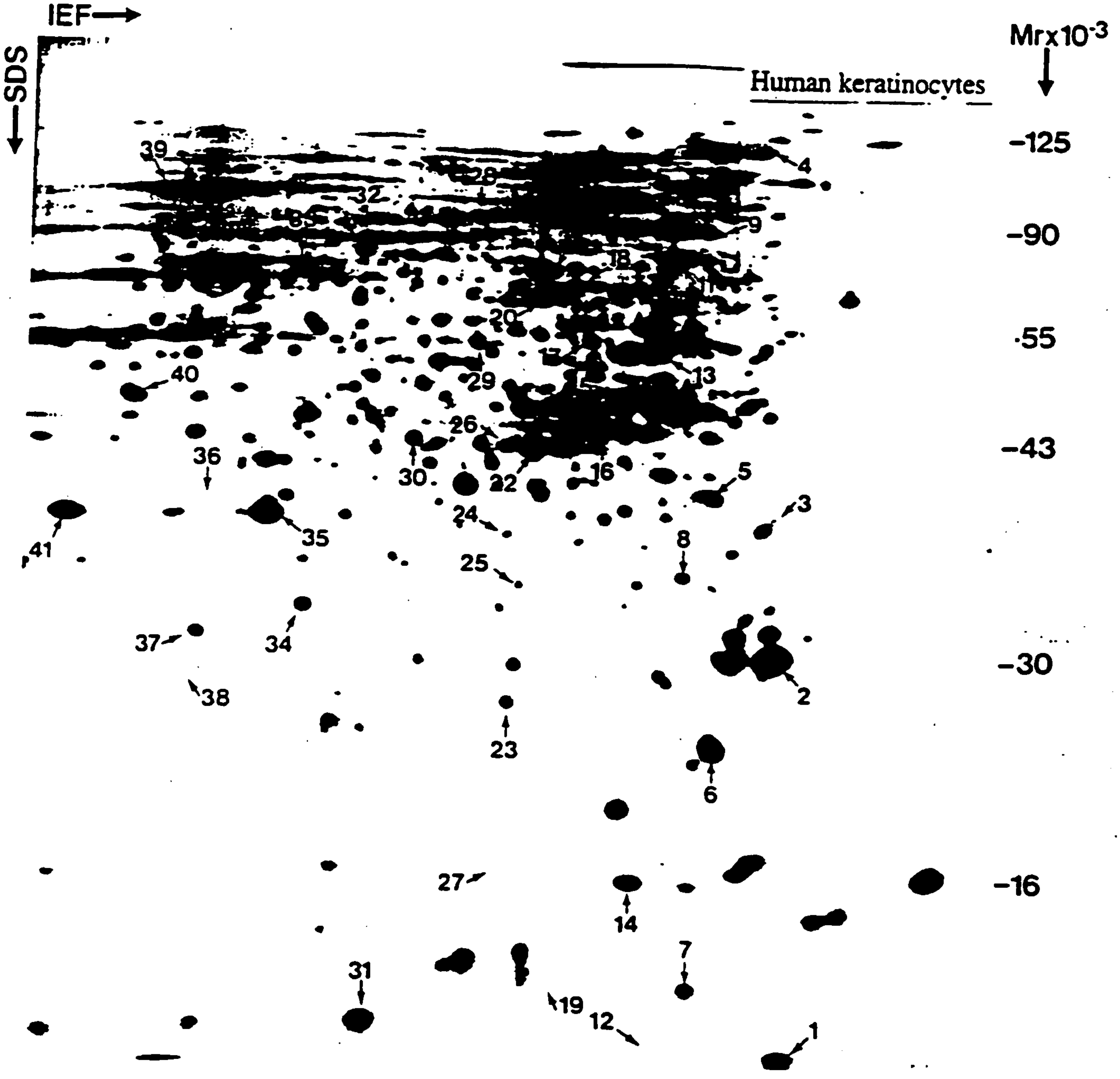


Figure 2. 2-D gel protein map of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human keratinocytes focused with CA-IEF in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

Table 2. Proteins from the human keratinocyte database localized in 2-D gels run with IPGs as first dimension

Number in Figs. 1-3	Protein name	IEF SSP number[a)]	Experimental pI value	Calculated pI value	Discrepancy (pH units)	Calculated net charge at experimental pI value	Buffer capacity charge units pro pH unit	N-terminal	Recalculated for suspected blockage		N-terminal	Swiss-Prot accession number
									pI value	Discrepancy pH units	Net charge	
1	CaN 19	9027	4.46		–							
2	Stratilin, bovine 14-3-3 related protein	9109	4.58									
3	Proliferating nuclear antigen (PCNA)/cyclin	9226	4.58	4.57	-0.01	-0.1	20.8	M				P12004
4	Involucrin	9703	4.63	4.63	0.00	-0.3	70.1	M				P07476
5	Nucleolar protein B23	8207	4.75	4.64	-0.11	-3.2	30.4	M				P06748
6	Translationally controlled tumor protein	8114	4.79	4.84	0.05	0.6	13.1	M[b)]				P13693
7	Thioredoxin	8006	4.86	4.82	-0.04	-0.3	7.1	V[b)]				P10599
8	Annexin V	8213	4.89	4.88	-0.01	-0.1	20.3	A[c)]				P08758
9	Heat shock protein 90-β	8611	4.95	4.94	-0.01	-0.5	56.2	P				P07900
10	Heat shock protein 90-α	2629	4.97	4.97	0.00	0.2	53.6	P				P08238
11	Glucose regulated protein 78 (BiP)	8515	4.99	4.98	-0.01	-0.6	37.5	E				P11021
12	Calcyclin	8017	5.02	5.32	0.30	1.3	3.6	M	5.09	0.07	0.3	P06703
13	Vimentin	8417	5.05	5.06	0.01	0.2	27.1	S				P08670
14	Initiation factor 4D	8016	5.05	5.08	0.03	0.2	7.6	A[c)]				P10159
15	Keratin 14	7305	5.08	5.09	0.01	0.2	21.0	T				P02533
16	β-Actin	7316	5.21	5.21	0.00	0.06	13.3	D[c)]				P02570
17	Heat shock protein 60	6403	5.23	5.24	0.01	0.1	17.5	A[b)]				P10809
18	Heat shock cognate 71KD	6501	5.28	5.37	0.09	1.8	18.1	M	5.32	0.04	0.8	P11142
19	Cystatin	6011	5.30	5.38	0.08	0.2	3.0	M				P01040
20	L-plastin	6412	5.34	5.41	0.07	1.1	17.7	M	5.36	0.02	0.3	P13797
21	Calelectrin	5628	5.35	5.37	0.02	0.5	21.3	M				P08133
22	Plasminogen activator inhibitor-2	6314	5.38	5.46	0.08	0.9	10.7	A[c)]	5.37	0.01	0.07	P05120
23	Glutathione S-transferase π	5101	5.43	5.44	0.01	0.08	3.9	M				P09211
24	Annexin VIII	5213	5.45	5.56	0.11	1.0	8.7	P	5.46	0.01	0.05	P13928
25	Annexin III	5204	5.46	5.63	0.17	1.4	8.4	M	5.52	0.06	0.5	P12429
26	Adenosine deaminase	5305	5.47	5.63	0.16	1.8	10.8	M	5.54	0.07	0.8	P00813
27	Stathmin	5001	5.55	5.61	0.06	0.4	6.6	M				P16949
28	Gelsolin, cytoplasmic	5608	5.59	5.58	-0.01	-0.1	16.5	A[c)] V				P06396
29	Rat phosphoinositide specific protein homolog	5410	5.62		–							
30	Elastase inhibitor	4114	5.74		–							
31	S100, calgizarin	4006	5.75		–							
32	Cytvillin, ezrin	3504	5.99	5.95	-0.04	-0.5	13.2	P				P15311
33	Moesin	3515	6.11	6.09	-0.02	-0.2	9.8	P				P26038
34	Purine nucleoside phosphorylase	2108	6.11	6.45	0.34	1.8	4.4	M	6.28	0.17	0.9	P00491
35	Annexin I	2216	6.18	6.64*	0.46	1.6	2.5	A	6.33	0.15	0.6	P04083
36	Aldose reductase	1202	6.40	6.55	0.15	0.7	4.2	A	6.41	0.01	0.2	P15121
37	Phosphoglycerate mutase (B form)	1107	6.46	6.75	0.29	0.9	2.6	A	6.46	0.00	0.0	P18669
38	Triosephosphate isomerase	1111	6.53	6.51	-0.02	-0.04	2.3	A[b)]				P00938
39	Elongation factor 2	1610	6.43	6.38	-0.05	-0.5	9.8	M				P13639
40	α-Enolase	1325	6.62	6.99	0.37	1.0	2.2	S	6.75	0.13	0.1	P06733
41	Annexin II	210	7.30	7.36	0.06	0.05	0.9	S[c)]				P07355

a) SSP number in the keratinocyte database [15]
b) Peptides N-terminally sequenced as liver proteins [3]
c) Peptides given as N-terminally blocked in Swiss-Prot database

3.2 Comparison between the determined and calculated p*I* values for human keratinocyte proteins

Thirty six of the 41 proteins listed in Table 2 are found in the Swiss-Prot database. Contrary to the plasma and liver proteins used in [9], the p*I* calcuations on the proteins used in this study posed some problems that reflected the way in which they were characterized. The proteins used by Bjellqvist *et al.* [9] were either very abundant and well-characterized plasma proteins or they were identified by *N*-terminal sequencing and, therefore, the nature of the *N*-terminals (acetylated or non-acetylated) was in both cases known. The proteins used in this study have all been characterized by internal sequencing [7] and it is known that *N*-terminal acetylation occurs with high frequency in eukaryotes.

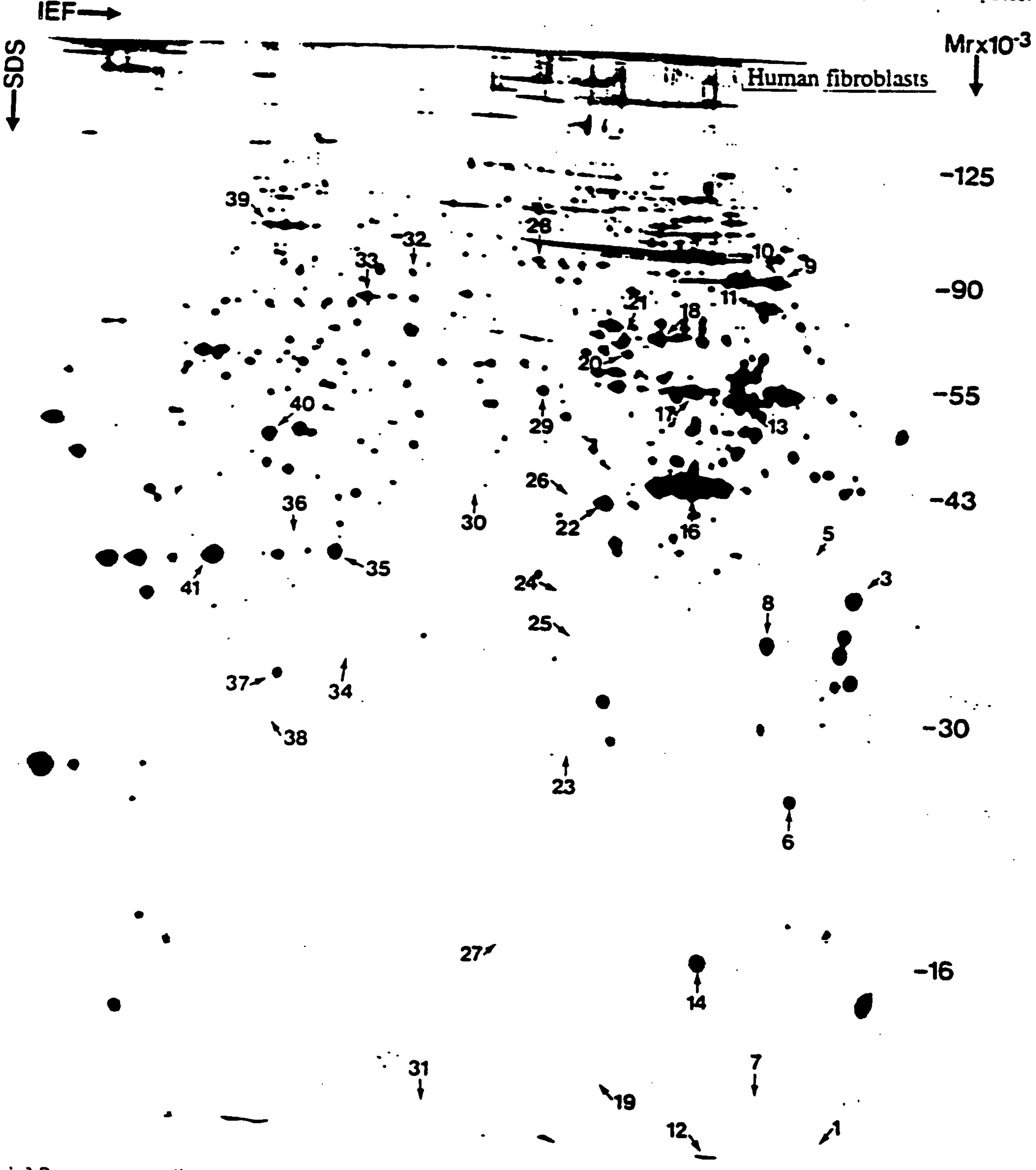


Figure 3. 2-D protein map of [^{35}S]methionine-labeled proteins from normal human fibroblasts focused with the nonlinear, wide-range IPG in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

According to Brown and Robert [25], proteins with acetylated N-terminals correspond in weight to approximately 80% of the soluble protein in ascites cells. Based on results from N-terminal sequencing, at least 40% of the spots in the human liver protein 2-D gel map appear to be blocked [3]. The corresponding number, derived from 107 spots in the 2-D gel map of human T-lymphocyte proteins, falls between 60 and 65% (J. Strahler, personal communication). Information concerning N-terminal blockage is not normally available, and in the Swiss-Prot database only 6 of the 36 keratinocyte proteins are specified as N-terminally blocked. We have, within the present material, defined 18 proteins for which the N-terminals are very likely to be correctly described. Six of these proteins are listed in the Swiss-Prot database as N-terminally blocked, four represent proteins which appear in the human liver 2-D gel map and have been N-terminally sequenced as liver proteins [3] and the remaining eight have N-terminal groups other than M, S and A, *i.e.* N-terminals for which N-acetylation is uncommon [26]. In Figs. 4A, B, C and D p*I* values calculated from Swiss Prot database information are plotted against the experimentally determined p*I* values for all the keratinocyte proteins listed in Table 2 and for the 18 selected proteins, as well as for the plasma and liver proteins (data from [9] valid for 10°C)*.

The calculations show that without knowledge of the status of the N-terminal group, precise predictions of p*I* values for eukaryotic proteins cannot be achieved based on the information available in Swiss-Prot and similar databases. However, for proteins where the N-terminal status is known, we find good correlation between predicted and experimental p*I* values. When the variance of the p*I* discrepancies and the variance of calculated charges at the experimental p*I* values derived from the present data set are compared with the corresponding

* There are four plots: (A) the 36 polypeptides from normal human keratinocytes (no corrections), (B) the 36 polypeptides from Fig. 4A where p*I* values have been recalculated for 12 polypeptides with M, S and A as N-terminally assumed blocked, based on calculated charge, (C) the 18 selected polypeptides with information on the N-terminal configuration, and (D) plasma and liver proteins.

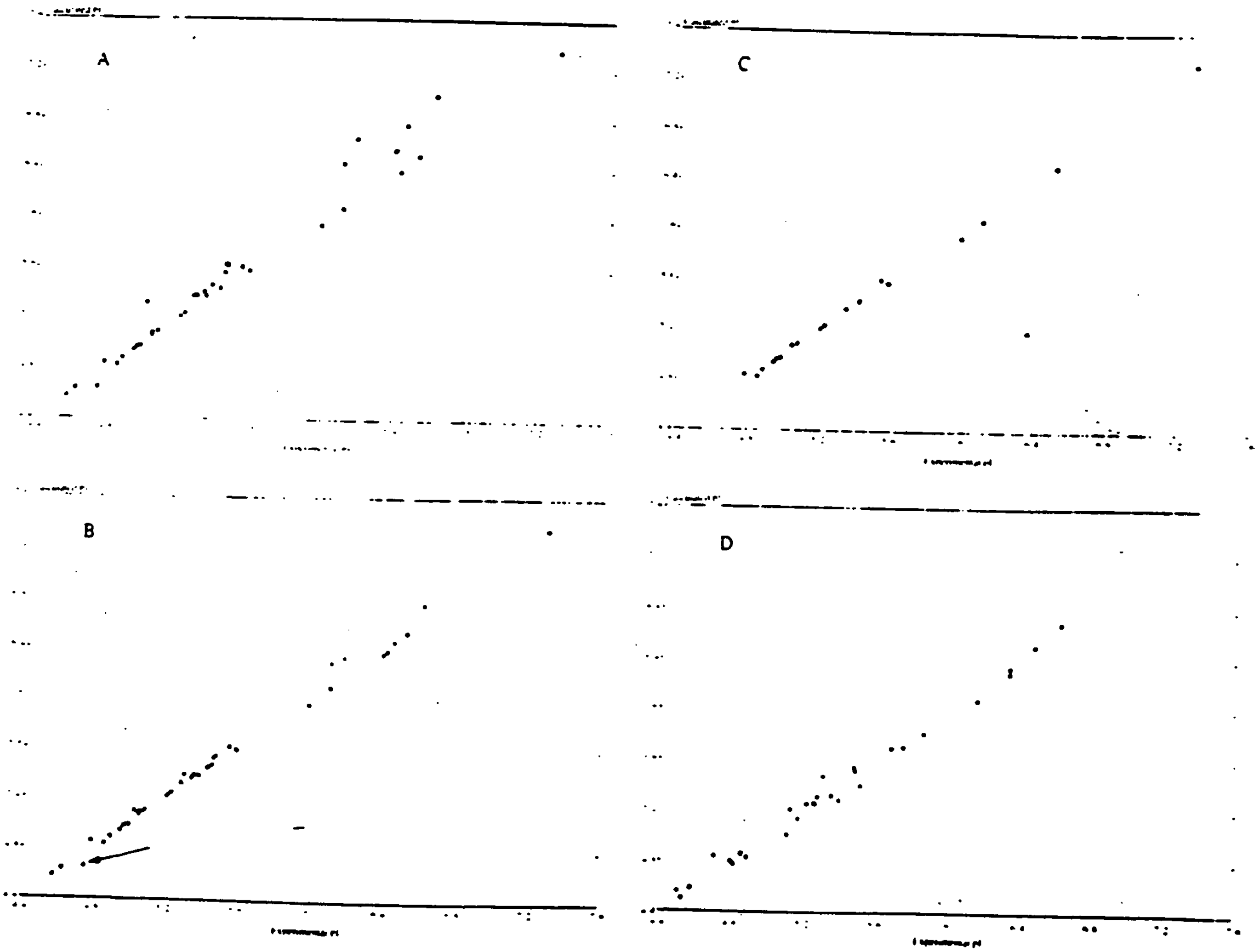


Figure 4. Calculated *vs.* experimental p*I* values. Lines are fitted using the least squares criterion. (A) 36 polypeptides from normal human keratinocytes (no corrections). (B) 36 polypeptides from Fig. 4A (including the 18 marker polypeptides) where p*I* values have been recalculated assuming N-terminal blockage; x indicates recalculated p*I* values; nucleolar protein B23 is indicated with an arrow. (C) 18 polypeptides with information on N-terminal configuration and (D) plasma and liver proteins.

values derived from the data on plasma and liver proteins in [9] (Table 3), the present data are found to result in larger variances for the values of both p*I* discrepancies and calculated charge at the experimental p*I* value when no information on posttranslational modification is taken into consideration. Correction for possible *N*-acetylation of 12 polypeptides with M, S and A as *N*-terminal results in a smaller variance of p*I* discrepancies, although not significantly different from values derived from [9], whereas the variance of the calculated charge at the experimental p*I* value is significantly higher. For the 18 selected proteins the variance for the p*I* discrepancies is significantly smaller than for the data in [9]; however, the corresponding value for calculated charge at the experimental p*I* value does not improve to the same extent. This, we believe, reflects another difference between the two sets of proteins used for the calculations. Based on spot distributions in 2-D gel maps, the set of proteins used here has a molecular weight distribution that is more representative of the patterns observed in mammalian cells. In the study by Bjellqvist *et al.* [9] most of the high molecular weight plasma proteins had to be excluded due to their unknown content of sialic acid which made the proteins analyzed in this study heavily biased towards low molecular weight proteins. The buffer capacity of proteins normally increases with the protein's molecular weight, and the average buffer capacity of the presently selected proteins with assumed known *N*-terminals is 18 charge units/pH unit, while the corresponding value for the proteins used in [9] is only 9 charge units/pH unit. High buffer capacity can be expected to improve the agreement between calculated and experimental p*I* values. Inspection of the data presented in Table 2 for the polypeptides with assumed known *N*-terminals verifies the importance of the buffer capacity. For 8 polypeptides having buffer capacities higher than 15 charge units/pH unit, the calculations in all cases yielded p*I* discrepancies with absolute values of less than 0.02 pH units. The largest discrepancy, 0.06 pH units, was observed for annexin II and stathmin, proteins which have low buffer capacity: 0.9 and 6.6 charge units/pH unit, respectively. The probability that the focusing position of a protein with known composition will fall within a certain distance from the calculated p*I* value therefore cannot be predicted by the variance alone. The buffer capacity of the specific protein must be taken into consideration as well. As indicated by the decrease of the variance of calculated charges at the experimental p*I* value for the selected proteins, the observed improvement can not solely be due to the higher buffer capacity of the keratinocyte proteins. The two studies relate to different experimental conditions. Good agreement between experimental and calculated p*I* values implies that the proteins are defolded and a factor that may contribute to the observed improvement is a more complete defolding of proteins caused by the higher temperature and urea concentration used in this study.

The data indicated that the precision with which p*I* values can be predicted for polypeptides with high buffer capacity is better than the precision with which experimental p*I* values can be determined. If the pH is defined through the p*K* values of the immobilized groups in the IPG containing gel, the precision of the experimentally calculated data will depend on the pH difference between the p*I* and the p*K* value of the immobilized group with the closest p*K*. For the present study this will give p*I* determinations with a precision varying in the range of ± 0.02–0.05 pH units [9]. The good agreement observed between the calculated and experimental p*I* values is due to the fact that errors are mainly systematic and, as discussed in [9], they will largely be cancelled out in the calculations. A pH scale defined through the presently determined p*I* values will not necessarily reflect the variation of the hydrogen ion activity during the focusing step in an optimal way, but it still allows precise predictions of focusing positions for polypeptides with known compositions, including information on posttranslational modifications. Calculated net charge at the experimentally found isoelectric point defined in this scale will serve as a tool to verify that the polypeptide

Table 3. Mean values and variances for the difference (experimental p*I*-calculated p*I*) in pH units and calculated charges at the experimental p*I* values, respectively

	Plasma and liver proteins (8 M urea, 10°C)		Keratinocyte proteins (9.8 M urea, 25°C)					
			All peptides		All peptides after correction for *N*-acetylation		Known *N*-terminal configuration (or very likely configuration)	
Number of proteins	29		36		36		18	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
Experimental p*I*-calculated p*I*	–0.011	0.005	0.072	0.017	0.019	0.003	0.005	0.001
F-value (p*I* discrepancy)[a)]	1		3.4		1.67		5	
P-level (p*I* discrepancy)[b)]	0.5		0.0005		0.0721		0.0004	
Calculated charge at the experimental p*I* value	–0.070	0.227	0.321	0.871	0.009	0.444	–0.014	0.109
F-value (calculated charge at the experimental p*I* value)[a)]	1		3.8		1.96		2.08	
P-level (calculated charge at the experimental p*I* value)[b)]	0.5		0.0002		0.0338		0.0536	

[a)] Comparison to the data in [9]. $F = S_1^2/S_2^2$, where S_1^2 is the larger of the two variances

[b)] $P(F(\nu_1, \nu_2) \geq F\text{-value})$, where ν_1 and ν_2 are the degrees of freedom for s_1 and s_2, respectively composition used in the calculation is correct and complete. Exceptions to this are proteins such as involucrin and heat shock protein 90 that have very high buffer capacities. Introduction of an extra charge unit into these proteins will only result in p*I* shifts falling in the range of 0.01—0.02 pH units and the effect is that the quality of the pH definition — the precision by which p*K* values used in the calculations are given and the precision of experimental p*I* values in these cases — will limit the possibilities to verify polypeptide compostion based on the experimental p*I* value.

Statistical comparison of experimental and calculated p*I* values was done using the *t*-test for dependent samples and normality of the discrepancies was estimated by probability plots. For the 36 proteins. the *p*-level is 0.0021. indicating that a result like this is unlikely to be a chance effect and must be assumed to represent a real difference. After correction for the most likely *N*-terminal configuration. the *p*-level is 0.043 and cannot be accepted as representing the same population since the *p*-level is less than 0.05 — the traditional *p*-limit of statistical significance. For the 18 proteins with a known or very likely *N*-terminal configuration the *t*-test gave a *p*-level of 0.49. which verifies that the experimental and calculated p*I* values are not significantly different.

Besides showing that p*I* values for denatured proteins with known compositions can be calculated with a high degree of precision from average p*K* values. the results also provide strong support for the notion that *N*-terminal blockage heavily depends on the nature of the *N*-terminal groups [26]. The results seem to indicate that with *N*-terminals other than M. S and A. only a few proteins have blocked *N*-terminals (1 out of 10 proteins in the present study). while it can be inferred from the data presented in Table 2 that a majority of the proteins with M. S and A as *N*-terminal are blocked. After correction for the effect of suspected *N*-terminal blockage there is only one protein (nucleolar protein B23) out of the 36 used in this study. which. in spite of a high buffer capacity. has a marked difference of 0.11 pH units between predicted and determined p*I* values (Fig. 4B): this corresponds to 3 charge units due to the high buffer capacity of this protein. This discrepancy in p*I* prediction and calculation of net charge at the p*I* is probably not due to deficiencies in the database information but instead reflects a shortcoming of the model used for p*I* calculations. Nucleolar protein B23 contains a domain extremely rich in aspartic and glutamic acid residues (Table 4). in which 26 out of 28 amino acid residues from position 161 to 188 are either a D or an E. A calculation based on the use of average p*K* values uninfluenced by the charged neighboring amino acid residues cannot be expected to correctly describe the p*I* value with almost half of the acidic groups packed together into a highly negatively charged region. This limitation caused by calculations based on average p*K* values does not severely limit the usefulness of the approach since a search through Swiss-Prot shows that this type of D/E-rich motif is uncommon. and the existence of a highly charged region is immediately apparent upon inspection of the amino acid sequence.

Table 4. Amino acid sequence of nucleolar phosphoprotein B23

1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
51	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
101	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
151	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
201	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
251	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

The quality of the information available in databases. especially concerning posttranslational modifications. is a major problem when the data is to be used for p*I* predictions. The *p*-level of 0.043 found for all 36 proteins after correction for *N*-acetylation. shows that this problem is not only limited to *N*-terminal blockage and the very good agreement found for the eighteen polypeptides. with assumingly correctly described *N*-terminal (Fig. 4C). must be regarded as an exception from this point of view. *N*-Terminal blockage is generally the main problem in relation to p*I* predictions for eukaryotic proteins. Of the 36 keratinocyte proteins analyzed. 18—20 are suspected to be *N*-terminally blocked (6 proteins blocked according to Swiss-Prot. 12 proteins with M. S or A as *N*-terminal and assumingly blocked based on the calculated charge. and two proteins. involucrin and nucleolar protein B23. with M as *N*-terminal for which the data does not allow any conclusion). This is in reasonable agreement with the conclusions based on the *N*-terminal sequencing data derived in connection with 2-D gel electrophoresis. *N*-terminal blockage can be suspected for 17—19 of the 26 proteins with M. S or A as *N*-terminal. while only 1 in 10 proteins with other *N*-terminal groups are blocked. The information that the frequency of *N*-terminal blockage is strongly related to the nature of the *N*-terminal group will be of some help in connection with p*I* predictions based on database information. However. without information from other sources. an uncertainty will always remain as to whether the *N*-terminal charge should be included in the p*I* calculation.

4 Concluding remarks

The data presented here lays the foundation for comparing 2-D gel protein maps of different cell types generated with nonlinear. wide-range IPGs in the first dimension. The focusing positions of 41 polypeptides common to most human cell types have been described in a pH scale that allows focusing positions to be predicted with a high degree of accuracy. provided that the composition of the polypeptides are known and that information on posttranslational modifications are available. For polypeptides with a very high buffer capacity. the limiting factor is the precision with which experimental pH values can be determined rather than the precision of the calculations. Possible deficiencies in the pH scale description of the variation of the hydrogen ion activity has. at least at the present state. no consequences for its practical use. The major limitation in connection with predictions of focusing positions from polypeptide compositions is the quality of existing data on protein compositions. especially concerning posttranslational modifications. Amino acid sequences have been reasonably easy to obtain. while posttranslational modifications have been difficult and work-intensive to determine. Recent developments in the field of mass spectrometry are fast changing this situation and within the next years we can expect a surge in reliable data in this area. While awaiting this development, verification of correctness and completeness of available information on polypeptide composition can be provided by experimental p*I* values in a pH scale based on the p*I* values determined in this study. So far, our data cover the pH range below pH ≈ 7.5. The basic pH range covered by NEPHGE as first dimension will be covered in forthcoming work.

Received December 29, 1993

5 References

[1] Gianazza, E., Astrua-Testori, S., Caccia, P., Giacon, P., Quaglia, L., Righetti, P. G., *Electrophoresis* 1986, *7*, 76–83.

[2] Gorg, A., Postel, W., Gunther, S., *Electrophoresis* 1988, *9*, 531–546.

[3] Hochstrasser, D. F., Frutiger, S., Paquet, N., Bairoch, A., Ravier, F., Pasquali, C., Sanchez, J.-C., Tissot, J.-D., Bjellqvist, B., Vargas, R., Appel, R. D., Hughes, G. J., *Electrophoresis* 1992, *13*, 992–1001.

[4] *Immobiline DryStrip Kit for 2-D Electrophoresis: Instructions*, Pharmacia LKB Biotechnology AB, Uppsala 1993.

[5] Anderson, N. L., Hickman, B. J., *Anal. Biochem.* 1979, *93*, 312–320.

[6] Neidhardt, F. C., Appleby, D. A., Sankar, P., Hutton, M. E., Phillips, T. A., *Electrophoresis* 1989, *10*, 116–121.

[7] Rasmussen, H. H., Damme, J. V., Puype, M., Gesser, B., Celis, J. E., Vandekerckhove, J., *Electrophoresis* 1992, *13*, 960–969.

[8] Gianazza, E., Artoni, G., Righetti, P. G., *Electrophoresis* 1983, *4*, 321–326.

[9] Bjellqvist, B., Hughes, G. J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J.-C., Frutiger, S., Hochstrasser, D. F., *Electrophoresis* 1993, *14*, 1023–1031.

[10] Bjellqvist, B., Pasquali, C., Ravier, C., Sanchez, J.-C., Hochstrasser, D. F., *Electrophoresis* 1993, *14*, 1357–1365.

[11] O'Farrell, P. H., *J. Biol. Chem.* 1975, *250*, 4007–4021.

[12] Görg, A., *Biochem. Soc. Transactions* 1993, *21*, 130–132.

[13] Hanash, S. M., Strahler, J. R., Neel, J. V., Hailat, N., Melhem, R., Keim, D., Zhu, X. X., Wagner, D., Gage, D. A., Watson, J. T., *Proc. Natl. Acad. Sci. USA* 1991, *88*, 5709–5713.

[14] Görg, A., Postel, W., Friedrich, C., Kuick, R., Strahler, J. R., Hanash, S. M., *Electrophoresis* 1991, *12*, 653–658.

[15] Celis, J. E., Rasmussen, H. H., Olsen, E., Madsen, P., Leffers, H., Honore, B., Dejgaard, K., Gromov, P., Hoffmann, H. J., Nielsen, M., Vassilev, A., Vintermyr, O., Hao, J., Celis, A., Basse, B., Lauridsen, J. B., Ratz, G. P., Andersen, A. H., Walbum, E., Kjærgaard, I., Puype, M., Van Damme, J., Delay, B., Vandekerckhove, J., *Electrophoresis* 1993, *14*, 1091–1198.

[16] Celis, J. E., Madsen, P., Rasmussen, H. H., Leffers, H., Honore, B., Gesser, B., Dejgaard, K., Olsen, E., Magnusson, N., Kiil, J., Celis, A., Lauridsen, J. B., Basse, B., Ratz, G. P., Andersen, A., Walbum, E., Brandstrup, B., Pedersen, P. S., Brandt, N. J., Puype, M., Van Damme, J., Vandekerckhove, J., *Electrophoresis* 1991, *12*, 802–872.

[17] Bjellqvist, B., Ek, K., Righetti, P. G., Gianazza, E., Gorg, A., Postel, W., Westermeier, R., *J. Biochem. Biophys. Methods* 1982, *6*, 317–333.

[18] Bairoch, A., Boeckman, B., *Nucleic Acids Res.* 1991, *19*, 2247–2249.

[19] Honoré, B., Madsen, P., Basse, B., Andersen, A., Walbum, E., Celis, J. E., Leffers, H., *Nucleic Acids Res.* 1990, *18*, 6692.

[20] Altland, K., *Electrophoresis* 1990, *11*, 140–147.

[21] Perrin, D. D., Dempsey, B., Serjeant, E. P., *pKa Predictions for Organic Acids and Bases*, Chapman and Hall Ltd., London 1981.

[22] Perrin, D. D., *Dissociation Constants of Organic Bases in Aqueous Solutions*, Butterworths, London 1965.

[23] Perrin, D. D., *Dissociation Constants of Organic Bases in Aqueous Solutions*, Supplement 1972, Butterworths, London 1972.

[24] Altland, K., Becher, P., Rossman, U., Bjellqvist, B., *Electrophoresis* 1988, 9, 474–485.

[25] Brown, J. L., Robert, W. K., *J. Biol. Chem.* 1976, *251*, 1009–1014.

[26] Persson, B., Flinta, C., Heine, G., Jörnvall, H., *Eur. J. Biochem.* 1985, *152*, 523–527.

G

Bengt Bjellqvist*
Bodil Basse
Eydfinnur Olsen
Julio E. Celis

Institute of Medical Biochemistry and Danish Centre for Human Genome Research, Aarhus University, Aarhus

Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions

A highly reproducible, commercial and nonlinear, wide-range immobilized pH gradient (IPG) was used to generate two-dimensional (2-D) gel maps of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human epidermal keratinocytes. Forty one proteins, common to most human cell types and recorded in the human keratinocyte 2-D gel protein database were identified in the 2-D gel maps and their isoelectric points (p*I*) were determined using narrow-range IPGs. The latter established a pH scale that allowed comparisons between 2-D gel maps generated either with other IPGs in the first dimension or with different human protein samples. Of the 41 proteins identified, a subset of 18 was defined as suitable to evaluate the correlation between calculated and experimental p*I* values for polypeptides with known composition. The variance calculated for the discrepancies between calculated and experimental p*I* values for these proteins was 0.001 pH units. Comparison of the values by the *t*-test for dependent samples (paired test) gave a *p*-level of 0.49, indicating that there is no significant difference between the calculated and experimental p*I* values. The precision of the calculated values depended on the buffer capacity of the proteins, and on average, it improved with increased buffer capacity. As shown here, the widely available information on protein sequences cannot, *a priori*, be assumed to be sufficient for calculating p*I* values because post-translational modifications, in particular *N*-terminal blockage, pose a major problem. Of the 36 proteins analyzed in this study, 18–20 were found to be *N*-terminally blocked and of these only 6 were indicated as such in databases. The probability of *N*-terminal blockage depended on the nature of the *N*-terminal group. Twenty six of the proteins had either M, S or A as *N*-terminal amino acids and of these 17–19 were blocked. Only 1 in 10 proteins containing other *N*-terminal groups were blocked.

1 Introduction

As compared with carrier ampholyte isoelectric focusing (CA-IEF), the application of immobilized pH gradients (IPGs) in the first dimension in 2-D gel electrophoresis offers improved reproducibility [1] because the nature of the pH gradient makes the resulting focusing positions insensitive to the focusing time [2] and to the type of sample applied [3]. The recently introduced ready-made IPG strips [4] seem to be an ideal substitute for the carrier ampholyte gradients, which until now have been the most commonly used first dimensions in 2-D gel electrophoresis. The availability of standardized first dimensions opens the possibility of comparing 2-D gel maps of various cell types generated in different laboratories, provided that the focusing positions of a number of easily recognizable polypeptide spots common to the cell types in question are known. Even though this approach is limited to experiments performed with the same standardized IPG, the flexibility provided by IPGs allows the pH gradient to be adjusted to the requirements of a particular experiment.

Exchange and communication of 2-D gel protein data requires a pH scale that is independent of the particular IPG used and by which the results can be described. The introduction of carbamylation trains and the relation of focusing positions to the spots in these trains represented a step forward towards solving the reproducibility problem experienced with carrier ampholyte focusing [5]. Problems associated with the use of carbamylation trains were mainly due to lack of temperature control and to the use of nonequilibrium focusing conditions. Accordingly, the pattern variation involved not only the resulting pH gradients, but also the relative spot positions as related to each other and to spots in the carbamylation trains. Even though the question of reproducibility has, to a large extent, been solved, the carbamylation trains are still not ideal as markers because the spots in the trains do not represent defined entities but rather a large number of differently carbamylated peptides having close p*I* values. As a result, the spots are large and poorly defined as compared to the ordinary polypeptide spots in 2-D gel maps.

Correspondence: Professor J. E. Celis, Institute of Medical Biochemistry and Danish Centre for Human Genome Research, Aarhus University, DK-8000 Aarhus C, Denmark

Abbreviations: CA-IEF, carrier ampholyte-isoelectric focusing; SSP, sample spot number

* Present addess: Pharmacia Biotech AB, S-751 82 Uppsala, Sweden

0173-0835/94/0304-0529 $5.00+.25/0

Neidhardt *et al.* [6] defined the pH gradient in 2-D gel experiments by p*I* markers whose p*I* values were calculated from the amino acid composition. Focusing positions of other polypeptides could be predicted from their composition but the p*K* values needed for the p*I* calculations were unknown. Various groups employing this approach do not use the same pK values [6, 7] and therefore, the p*I* values derived in this way cannot be expected to describe the variation of the hydrogen ion activity. In spite of this fact, it is still possible to make approximate predictions of focusing positions because the p*K* values used to define the pH gradient are also used to calculate p*I* values and to predict the focusing positions. Errors in p*K* assignments are therefore compensated. A pH scale which corretly reflects the variation in hydrogen ion activity during focusing should improve the precision of the predictions, but this has never been implemented with CA-IEF focusing as a first dimension in 2-D gel electrophoresis. The main reason for this are the problems associated with pH measurements in focused gels containing high concentrations of urea.

IPGs can be described from the concentration variation of the immobilized groups, provided that the p*K* values of these groups are known for the conditions prevailing during focusing. To avoid measurements on gels, Gianazza *et al.* [8] suggested the use of p*K* values derived by addition of determined p*K* shifts. Recently, direct determinations of p*K* differences between immobilized groups in IPGs were made by determining p*I*-p*K* values in overlapping narrow-range IPGs [9, 10] and the results verified the applicability of the Gianazza approach. A description of the focusing results in a pH scale, which correctly describes the variation of the hydrogen ion activity for the focusing conditions used, not only allows the comparison of 2-D gel maps generated with different IPGs, but also opens the possibility for correlating the focusing position of a polypeptide with its composition [9]. Experiments by Bjellqvist *et al.* [9, 10] have implied that pH scales showing good correlation between calculated and experimental p*I* values can be derived for any of the conditions commonly used for focusing in connection with 2-D gel electrophoresis. These pH scales are then defined through the p*K* values of the immobilized groups in the IPG containing gel. To be useful for interlaboratory comparisons, however, the pH scale has to be defined through p*I* values of easily recognizable spots present in the 2-D gel map. So far, p*I* determinations in a useful pH scale, combined with determinations of pK values needed for p*I* calculations, have only been made for the pH range 4.5—6.5 at 10°C [9]. CA-IEF focusing as described by O'Farrell [11] does not control the temperature of the first dimension, which can be expected to be slightly above room temperature. With IPGs, the temperature commonly used is about 20°C [4, 12] or 25°C [13] and this is a critical parameter that needs to be controlled [14].

The present work was designed to compare 2-D gel maps of different cell types in a laboratory applying both CA-IEF and IPG focusing at a common temperature. To this end we have generated 2-D gel maps of proteins from noncultured, unfractionated normal human epidermal keratinocytes with IPG in the first dimension and a focusing temperature of 25°C. We have used commercial nonlinear, wide-range IPG strips which give 2-D gel maps that are closely similar to the ones resulting with the CA-IEF technique used to establish the human keratinocyte database [15]. As an initial step towards interlaboratory comparisons of results obtained with the nonlinear gradient as a first dimension we report here on the focusing positions of 41 known proteins that are common to most human cell types. The pH range covered corresponds to the range in classical CA-IEF 2-D gel electrophoresis and in order to use these proteins as internal standards for comparing 2-D gel maps generated with other IPGs we determined their p*I* values with narrow-range IPGs in the first dimension. We have compared the calculated *versus* experimental p*I* values and show that it is necessary to have further information (absence or presence and nature of posttranslational modifications), in addition to amino acid composition to be able to calculate p*I* values that correspond to the actual experimental values. The p*K* values used for the calculations are provided and the usefulness of p*I* prediction in relation to database information is discussed. Furthermore, we comment on the possibility of using experimentally determined p*I* values to verify the available database information on polypeptide composition.

2 Materials and methods

2.1 Apparatus and chemicals

Equipment for isoelectric focusing and horizontal SDS electrophoresis (Multiphor® II electrophoresis chamber, Immobiline® strip tray, Multidrive XL programmable power supply, Macrodrive power supply and Multitemp® II) was from Pharmacia LKB Biotechnology AB (Uppsala, Sweden). Vertical second-dimensional gels were run in the home-made equipment described in [15]. The IPG strips with the wide-range nonlinear pH gradient were either Immobiline DryStrip® pH 3—10 NL, 180 mm or alternatively 160 mm long IPG strips with a corresponding pH gradient. In both cases the IPG strips were delivered by Pharmacia LKB. Immobiline, Pharmalyte, Ampholine, GelBond as well as PAG film and the ready-made horizontal SDS gels (ExcelGel® XL SDS 12—14) were also from Pharmacia LKB. Purified proteins and peptides were from Sigma (St. Louis, MO).

2.2 Sample preparation

Preparation and labeling of unfractionated keratinocytes as well as fibroblasts have been described in [16]. Cells were lysed in a solution containing 9.8 M urea, 2% w/v NP-40, 100 mM DTT and 2% v/v Ampholine pH 7—9.

2.3 2-D gel electrophoresis

First-dimensional focusing was performed according to Görg *et al.* [2] with some minor modifications, as described in [9]. Rehydration of the IPG strips was made in a solution containing 9.8 M urea, 2% w/v CHAPS, 10 mM DTT and 2% v/v carrier ampholyte mixture. The carrier ampholyte mixture consisted of 2 parts Pharmalyte

4–6.5, 1 part Ampholine pH 6–8 and 1 part Pharmalyte pH 8–10.5. Usually, cathodic sample application was used and the samples were diluted 2–20 times in a solution containing 9.8 M urea, 4% w/v CHAPS, 1% w/v DTT and 35 mM Tris base. For acidic application, the Tris-base was substituted with 100 mM acetic acid. The degree of dilution and sample volume (20–100 µL) depended on the particular sample and the IPG, and whether visualization of the proteins was to be done by Coomassie Brilliant Blue or silver staining. With the wide-range non-linear IPG, 10–30 µg of total protein was loaded for silver staining and 100–200 µg for Coomassie staining. Focusing was done overnight with Vh products in the range of 45–60 kVh with 160 mm long strips and 50–70 kVh with 180 mm long strips. Solubilization of polypeptides and blocking of -SH groups prior to the second-dimensional run, as well as loading on the second-dimensional gel was done as described in [9]. The stacking gel was omitted and 5–10 mm were left at the top of the second-dimensional gel for applying the IPG strip. The space was filled with electrode buffer containing 0.5% w/v agarose. Casting, running, staining and autoradiography were carried out as described in [15].

2.4 Experimental determination of p*I* values

The determination of the pK' differences between Immobilines pK' 4.6, pK' 6.2 and pK' 7.0 necessary for the calibration of the pH scale at 25 °C in 9.8 M urea was done as described in [9] with the same narrow-range IPGs. The pH scale was defined by setting the pK' value of Immobiline pK' 4.6 equal to 4.61 [9] and the determined pK' differences gave the pK' values of Immobilines pK' 6.2 and pK' 7.0, equal to 5.73 and 6.54, respectively. The pK' differences found are in good agreement with values derived from [17] and [8] by extrapolation to 9.8 M urea concentration. As in [9], additional narrow-range recipes have been used for determining p*I* values. With narrow-range IPGs extending to pH values higher than the pK' value of Immobiline pK' 7.0, anodic sample application was used with acetic acid added to the sample solution. Otherwise, cathodic sample application was used with the same sample buffer as for wide-range IPGs.

2.5 Protein compositions used for p*I* calculations

With the exception of vimentin, protein compositions are from the Swiss-Prot database [18]. For vimentin, we used the data from [19], where the amino acid at position 41 is a D instead of a S. Information in the Swiss-Prot database on phosphorylation has been disregarded because it was known from earlier studies (J. E. Celis, unpublished results) that the spots in question corresponded to the unphosphorylated forms of the peptides.

2.6 Calculation of p*I* values

For the p*I* calculations it was assumed that the same pK' value could be used for an amino acid residue in all polypeptides and in all positions in the peptide except for *N*- or *C*-terminally placed amino acids. For the pK' values of the *N*-terminal amino groups the effect of the different substituents on the α-carbon were taken into account. The calculations of p*I* values were made with the aid of the IPG-maker program [20].

2.7 pK' values used for p*I* calculations

For the carboxyl terminal group and internal glutamyl and aspartyl residues the same pK' values were used as in [9]. For *C*-terminal glutamyl and aspartyl residues, separate pK' values were derived with the aid of the Taft equations [9, 21]. The pK' values of histidyl groups were calculated from the p*I* values of human carbonic anhydrase I as in [9]. For *N*-terminal glycine a pK' value of 7.50 was used. The pK' shift caused by a substituent on the α-carbon was assumed to be identical with the pK' shift the substituent caused for the amino group in the amino acid, *i.e.* 2.28 pH units were subtracted from the pK' values for the amino groups in the amino acids given in [22, 23]. The approximate pK' value of 9 for the cystenyl group was taken from [24]. For tyrosyl and arginyl groups we used the pK' values for the amino acids [22, 23]. For lysyl groups the effect of high urea concentration on amino groups was taken into account and 0.5 pH units were subtracted from the amino acid pK' value. These last three pK' values are far from the pH range under study and the results found would have been the same if lysyl and arginyl groups were assumed to be fully ionized while the ionization of tyrosyl groups were neglected. A complete list of the pK' values used is given in Table 1.

Table 1. pK' Values used for the ionizable groups in peptides 9.8 M urea, 25 °C

Ionizable group	pK'
C-terminal	3.55*
N-terminal	
Ala	7.50
Met	7.00
Ser	6.93
Pro	8.36
Thr	6.82
Val	7.44
Glu	7.70
Internal	
Asp	4.05
Glu	4.45
His	5.98
Cys	9
Tyr	10
Lys	10
Arg	12
C-terminal side chain groups	
Asp	4.55
Glu	4.75

2.8 Statistical analysis

Statistical comparisons of the experimental and calculated p*I* values were done on an Apple Macintosh IIsi using the statistical package Statistica/Mac, release 3.0b (from StatSoft Inc., Tulsa, Oklahoma). Calculated and experimental p*I* values were compared by the *t*-test for correlated samples (paired *t*-test). The normality of p*I* differences was estimated graphically by probability plots. The variances of the data presented here and the similar data on plasma and liver proteins in [9] were compared by the F-test.

3 Results and discussion

3.1 Identification of polypeptides and p*I* determinations

The 2-D gel maps of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human keratinocytes, focused with the nonlinear, wide-range IPG and CA-IEF pH gradients in the first dimension, are shown in Figs. 1 and 2, respectively. The IPG extends to higher pH values but otherwise the two patterns are very similar and most of the spots in the IPG pattern can be directly related to the corresponding spots in the CA-IEF gel. To obtain comparable patterns it was important to keep the focusing temperature as similar as possible. Compared to other studies [1–4, 9, 10, 12–14], we increased the urea concentration in the focusing gel to 9.8 M because keratins streaked badly in the focusing dimension when 8 M urea was used, presumably due to

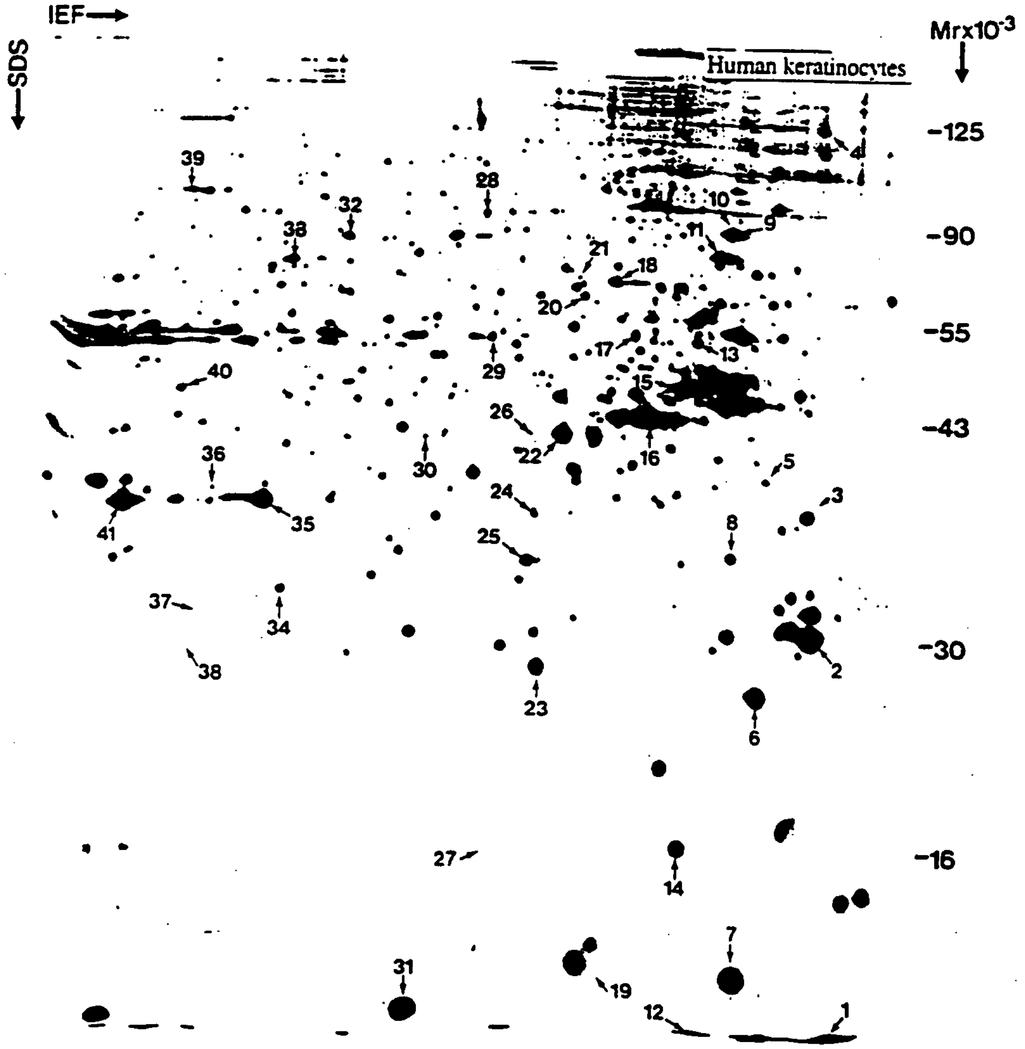


Figure 1. 2-D gel protein map of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human keratinocytes focused with the nonlinear, wide-range IPG in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

aggregates of acidic and basic keratins. An increase in urea concentration to 9 M or more eliminated these streaks: apart from this effect, no other major changes in the focusing positions were observed. In Fig. 1 we have indicated the positions of 41 known proteins from the human keratinocyte 2-D gel database that are most likely common to most human cell types. The choice was made because these proteins are easy to identify with certainty. With the exception of stratifin (spot 2), involucrin (spot 4) and keratin 14 (spot 15), which are all epithelial markers, these proteins are also present in human fibroblasts (Fig. 3) and lymphocytes (results not shown), and therefore can be used as landmarks for comparing 2-D gel maps derived from different cell types. In Table 2 the 41 proteins are listed together with their sample spot numbers (SSP) in the human keratinocyte protein database and p/ values determined in 2-D gel maps generated with narrow-range IPGs in the first dimension.

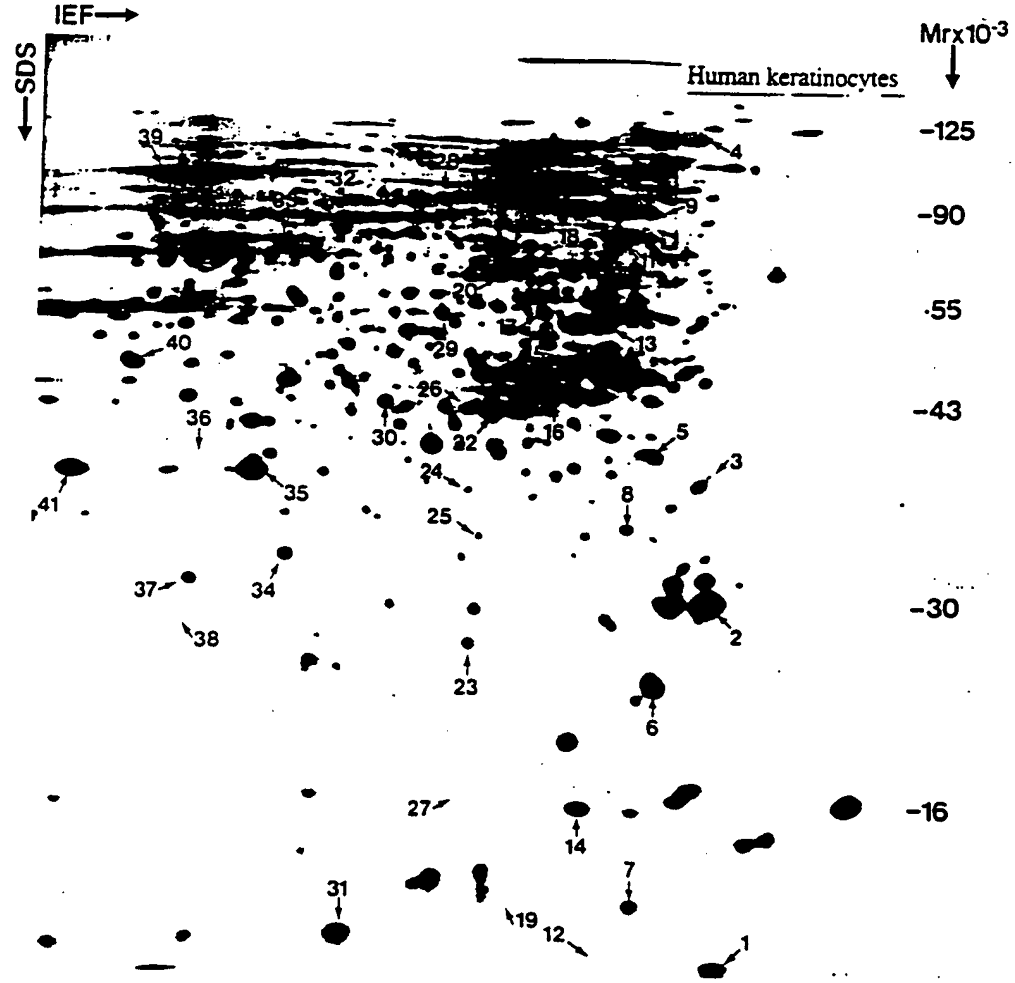


Figure 2. 2-D gel protein map of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human keratinocytes focused with CA-IEF in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

Table 2. Proteins from the human keratinocyte database localized in 2-D gels run with IPGs as first dimension

Number in Figs. 1-3	Protein name	IEF SSP number[a)]	Experimental pI value	Calculated pI value	Discrepancy (pH units)	Calculated net charge at experimental pI value	Buffer capacity charge units pro pH unit	N-terminal	Recalculated for suspected blockage		N-terminal	Swiss-Prot accession number
									pI value	Discrepancy pH units	Net charge	
1	CaN 19	9027	4.36									
2	Stratilin, bovine 14-3-3 related protein	9109	4.58		--							
3	Proliferating nuclear antigen (PCNA)/cyclin	9226	4.58	4.57	-0.01	-0.4	20.8	M				P12004
4	Involucrin	9703	4.63	4.63	0.00	-0.3	70.1	M				P07476
5	Nucleolar protein B23	8207	4.75	4.64	-0.11	-3.2	30.4	M				P06748
6	Translationally controlled tumor protein	8114	4.79	4.84	0.05	0.6	13.1	M[b)]				P13693
7	Thioredoxin	8006	4.86	4.82	-0.04	-0.3	7.1	V[b)]				P10599
8	Annexin V	8213	4.89	4.88	-0.01	-0.1	20.3	A[c)]				P08758
9	Heat shock protein 90-β	8611	4.95	4.94	-0.01	-0.5	56.2	P				P07900
10	Heat shock protein 90-α	2629	4.97	4.97	0.00	0.2	53.6	P				P08238
11	Glucose regulated protein 78 (BiP)	8515	4.99	4.98	-0.01	-0.6	37.5	E				P11021
12	Calcyclin	8017	5.02	5.32	0.30	1.3	3.6	M	5.09	0.07	0.3	P06703
13	Vimentin	8417	5.05	5.06	0.01	0.2	27.1	S				P08670
14	Initiation factor 4D	8016	5.05	5.08	0.03	0.2	7.6	A[c)]				P10159
15	Keratin 14	7305	5.08	5.09	0.01	0.2	21.0	T				P02533
16	β-Actin	7316	5.21	5.21	0.00	0.06	13.3	D[c)]				P02570
17	Heat shock protein 60	6101	5.21	5.24	0.03	0.1	17.5	A[b)]				P10809
18	Heat shock cognate 71KD	6501	5.28	5.37	0.09	1.8	18.1	M	5.32	0.04	0.8	P11142
19	Cystatin	6011	5.30	5.38	0.08	0.2	3.0	M				P01040
20	L-plastin	6112	5.34	5.41	0.07	1.3	17.7	M	5.36	0.02	0.3	P13797
21	Calelectrin	5628	5.35	5.37	0.02	0.5	21.3	A[c)]				P08133
22	Plasminogen activator inhibitor-2	6111	5.38	5.46	0.08	0.9	10.7	M	5.37	0.01	0.07	P05120
23	Glutathione S-transferase π	5101	5.43	5.44	0.01	0.08	3.9	P				P09211
24	Annexin VIII	5213	5.45	5.56	0.11	1.0	8.7	M	5.46	0.01	0.05	P13928
25	Annexin III	5204	5.46	5.63	0.17	1.4	8.4	M	5.52	0.06	0.5	P12429
26	Adenosine deaminase	5305	5.47	5.63	0.16	1.8	10.8	M	5.54	0.07	0.8	P00813
27	Stathmin	5001	5.55	5.61	0.06	0.4	6.6	A[c)]				P16949
28	Gelsolin, cytoplasmic	5608	5.59	5.58	-0.01	-0.1	16.5	V				P06396
29	Rat phosphoinoside specific protein homolog	5410	5.62		--		..					
30	Elastase inhibitor	4114	5.74		..	--						
31	S100, calgizarin	4006	5.75	..	–		..					
32	Cytvillin, ezrin	3504	5.99	5.95	-0.04	0.5	13.2	P				P15311
33	Moesin	3515	6.11	6.09	-0.02	-0.2	9.8	P				P26038
34	Purine nucleoside phosphorylase	2108	6.11	6.45	0.34	1.8	4.4	M	6.28	0.17	0.9	P00491
35	Annexin I	2216	6.18	6.64*	0.46	1.6	2.5	A	6.33	0.15	0.6	P04083
36	Aldose reductase	1202	6.40	6.55	0.15	0.7	4.2	A	6.36	0.04	0.2	P15121
37	Phosphoglycerate mutase (B form)	1107	6.46	6.75	0.29	0.9	2.6	A	6.46	0.00	0.0	P18669
38	Triosephosphate isomerase	1111	6.53	6.51	-0.02	-0.04	2.3	A[b)]				P00938
39	Elongation factor 2	1610	6.43	6.38	-0.05	-0.5	9.8	M				P13639
40	α-Enolase	1325	6.62	6.99	0.37	1.0	2.2	S	6.75	0.13	0.1	P06733
41	Annexin II	210	7.30	7.36	0.06	0.05	0.9	S[c)]				P07355

a) SSP number in the keratinocyte database [15]

b) Peptides N-terminally sequenced as liver proteins [3]

c) Peptides given as N-terminally blocked in Swiss-Prot database

3.2 Comparison between the determined and calculated p*I* values for human keratinocyte proteins

Thirty six of the 41 proteins listed in Table 2 are found in the Swiss-Prot database. Contrary to the plasma and liver proteins used in [9], the p*I* calcuations on the proteins used in this study posed some problems that reflected the way in which they were characterized. The proteins used by Bjellqvist *et al.* [9] were either very abundant and well-characterized plasma proteins or they were identified by *N*-terminal sequencing and, therefore, the nature of the *N*-terminals (acetylated or non-acetylated) was in both cases known. The proteins used in this study have all been characterized by internal sequencing [7] and it is known that *N*-terminal acetylation occurs with high frequency in eukaryotes.

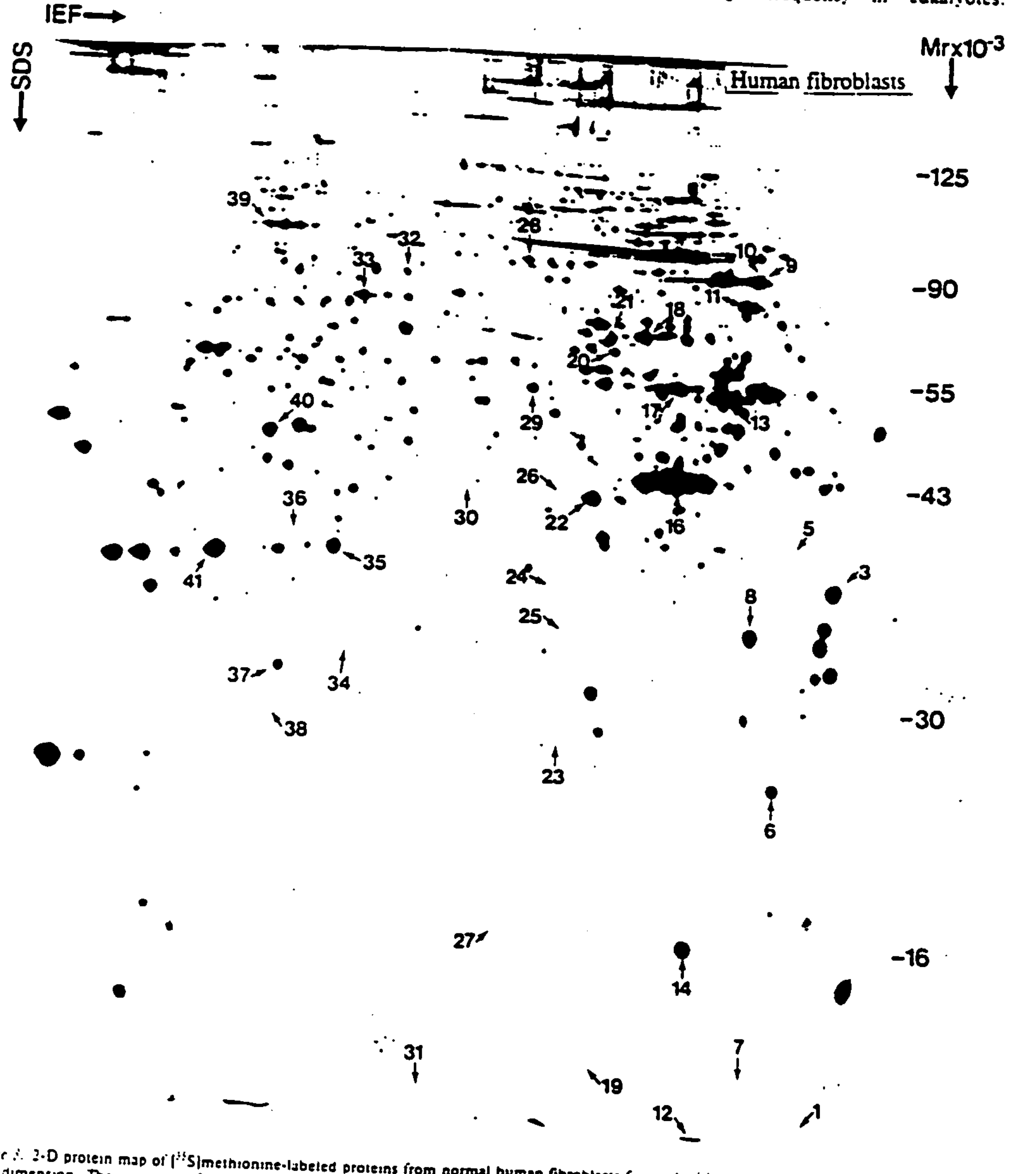


Figure 3. 2-D protein map of [^{35}S]methionine-labeled proteins from normal human fibroblasts focused with the nonlinear, wide-range IPG in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

According to Brown and Robert [25], proteins with acetylated *N*-terminals correspond in weight to approximately 80% of the soluble protein in ascites cells. Based on results from *N*-terminal sequencing, at least 40% of the spots in the human liver protein 2-D gel map appear to be blocked [3]. The corresponding number, derived from 107 spots in the 2-D gel map of human T-lymphocyte proteins, falls between 60 and 65% (J. Strahler, personal communication). Information concerning *N*-terminal blockage is not normally available, and in the Swiss-Prot database only 6 of the 36 keratinocyte proteins are specified as *N*-terminally blocked. We have, within the present material, defined 18 proteins for which the *N*-terminals are very likely to be correctly described. Six of these proteins are listed in the Swiss-Prot database as *N*-terminally blocked, four represent proteins which appear in the human liver 2-D gel map and have been *N*-terminally sequenced as liver proteins [3] and the remaining eight have *N*-terminal groups other than M, S and A, *i.e.* *N*-terminals for which *N*-acetylation is uncommon [26]. In Figs. 4A, B, C and D p*I* values calculated from Swiss Prot database information are plotted against the experimentally determined p*I* values for all the keratinocyte proteins listed in Table 2 and for the 18 selected proteins, as well as for the plasma and liver proteins (data from [9] valid for 10°C)*.

The calculations show that without knowledge of the status of the *N*-terminal group, precise predictions of p*I* values for eukaryotic proteins cannot be achieved based on the information available in Swiss-Prot and similar databases. However, for proteins where the *N*-terminal status is known, we find good correlation between predicted and experimental p*I* values. When the variance of the p*I* discrepancies and the variance of calculated charges at the experimental p*I* values derived from the present data set are compared with the corresponding

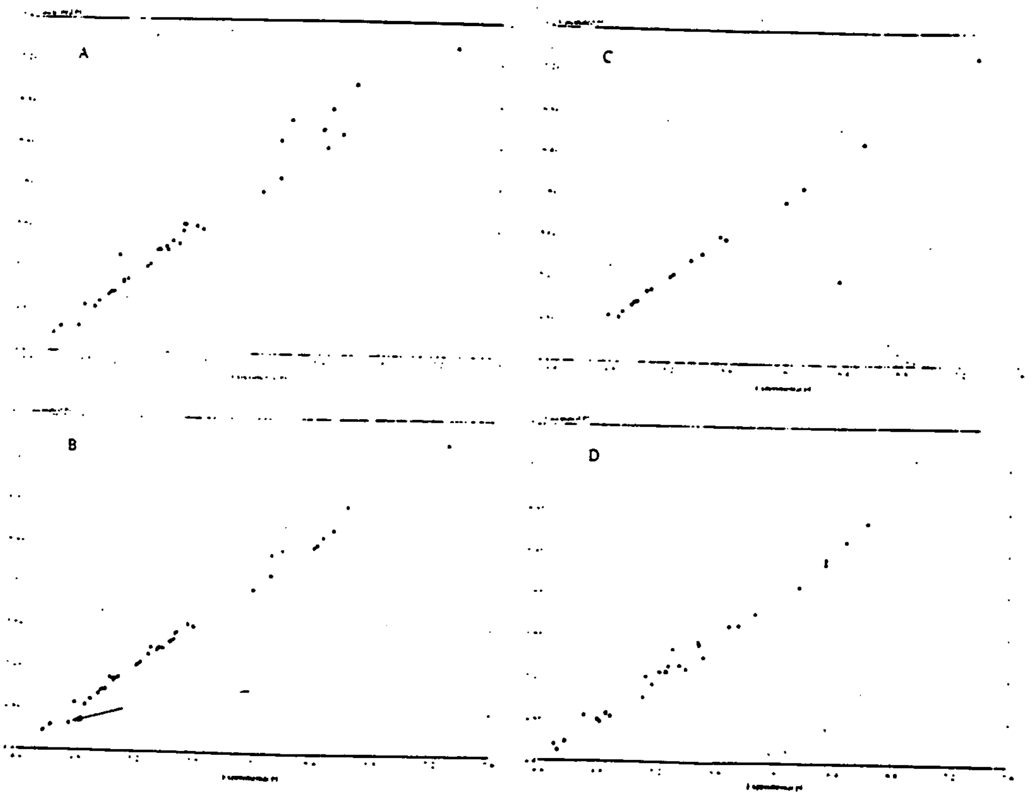


Figure 4. Calculated *vs.* experimental p*I* values. Lines are fitted using the least squares' criterion. (A) 36 polypeptides from normal human keratinocytes (no corrections). (B) 36 polypeptides from Fig. 4A (including the 18 marker polypeptides) where p*I* values have been recalculated assuming *N*-terminal blockage: x indicates recalculated p*I* values: nucleolar protein B23 is indicated with an arrow. (C) 18 polypeptides with information on *N*-terminal configuration and (D) plasma and liver proteins.

* There are four plots: (A) the 36 polypeptides from normal human keratinocytes (no corrections), (B) the 36 polypeptides from Fig. 4A where p*I* values have been recalculated for 12 polypeptides with M, S and A as *N*-terminally assumed blocked, based on calculated charge, (C) the 18 selected polypeptides with information on the *N*-terminal configuration, and (D) plasma and liver proteins.

values derived from the data on plasma and liver proteins in [9] (Table 3), the present data are found to result in larger variances for the values of both *pI* discrepancies and calculated charge at the experimental *pI* value when no information on posttranslational modification is taken into consideration. Correction for possible *N*-acetylation of 12 polypeptides with M, S and A as *N*-terminal results in a smaller variance of *pI* discrepancies, although not significantly different from values derived from [9], whereas the variance of the calculated charge at the experimental *pI* value is significantly higher. For the 18 selected proteins the variance for the *pI* discrepancies is significantly smaller than for the data in [9]; however, the corresponding value for calculated charge at the experimental *pI* value does not improve to the same extent. This, we believe, reflects another difference between the two sets of proteins used for the calculations. Based on spot distributions in 2-D gel maps, the set of proteins used here has a molecular weight distribution that is more representative of the patterns observed in mammalian cells. In the study by Bjellqvist *et al.* [9] most of the high molecular weight plasma proteins had to be excluded due to their unknown content of sialic acid which made the proteins analyzed in this study heavily biased towards low molecular weight proteins. The buffer capacity of proteins normally increases with the protein's molecular weight, and the average buffer capacity of the presently selected proteins with assumed known *N*-terminals is 18 charge units/pH unit, while the corresponding value for the proteins used in [9] is only 9 charge units/pH unit. High buffer capacity can be expected to improve the agreement between calculated and experimental *pI* values. Inspection of the data presented in Table 2 for the polypeptides with assumed known *N*-terminals verifies the importance of the buffer capacity. For 8 polypeptides having buffer capacities higher than 15 charge units/pH unit, the calculations in all cases yielded *pI* discrepancies with absolute values of less than 0.02 pH units. The largest discrepancy, 0.06 pH units, was observed for annexin II and stathmin, proteins which have low buffer capacity: 0.9 and 6.6 charge units/pH unit, respectively. The probability that the focusing position of a protein with known composition will fall within a certain distance from the calculated *pI* value therefore cannot be predicted by the variance alone. The buffer capacity of the specific protein must be taken into consideration as well. As indicated by the decrease of the variance of calculated charges at the experimental *pI* value for the selected proteins, the observed improvement can not solely be due to the higher buffer capacity of the keratinocyte proteins. The two studies relate to different experimental conditions. Good agreement between experimental and calculated *pI* values implies that the proteins are defolded and a factor that may contribute to the observed improvement is a more complete defolding of proteins caused by the higher temperature and urea concentration used in this study.

The data indicated that the precision with which *pI* values can be predicted for polypeptides with high buffer capacity is better than the precision with which experimental *pI* values can be determined. If the pH is defined through the p*K* values of the immobilized groups in the IPG containing gel, the precision of the experimentally calculated data will depend on the pH difference between the *pI* and the p*K* value of the immobilized group with the closest p*K*. For the present study this will give *pI* determinations with a precision varying in the range of ± 0.02–0.05 pH units [9]. The good agreement observed between the calculated and experimental *pI* values is due to the fact that errors are mainly systematic and, as discussed in [9], they will largely be cancelled out in the calculations. A pH scale defined through the presently determined *pI* values will not necessarily reflect the variation of the hydrogen ion activity during the focusing step in an optimal way, but it still allows precise predictions of focusing positions for polypeptides with known compositions, including information on posttranslational modifications. Calculated net charge at the experimentally found isoelectric point defined in this scale will serve as a tool to verify that the polypeptide

Table 3. Mean values and variances for the difference (experimental *pI*-calculated *pI*) in pH units and calculated charges at the experimental *pI* values, respectively

	Plasma and liver proteins (8 M urea, 10°C)		Keratinocyte proteins (9.8 M urea, 25°C)					
			All peptides		All peptides after correction for *N*-acetylation		Known *N*-terminal configuration (or very likely configuration)	
Number of proteins	29		36		36		18	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
Experimental *pI*-calculated *pI*	−0.011	0.005	0.072	0.017	0.019	0.003	0.005	0.001
F-value (*pI* discrepancy)[a]	1		3.4		1.67		5	
P-level (*pI* discrepancy)[b]	0.5		0.0005		0.0721		0.0004	
Calculated charge at the experimental *pI* value	−0.070	0.227	0.321	0.871	0.009	0.444	−0.014	0.109
F-value (calculated charge at the experimental *pI* value)[a]	1		3.8		1.96		2.08	
P-level (calculated charge at the experimental *pI* value)[b]	0.5		0.0002		0.0338		0.0536	

a) Comparison to the data in [9]. $F = S_1^2/S_2^2$, where S_1^2 is the larger of the two variances

b) $P(F(v_1, v_2) \geq F\text{-value})$, where v_1 and v_2 are the degrees of freedom for s_1 and s_2, respectively composition used in the calculation is correct and complete. Exceptions to this are proteins such as involucrin and heat shock protein 90 that have very high buffer capacities. Introduction of an extra charge unit into these proteins will only result in p*I* shifts falling in the range of 0.01–0.02 pH units and the effect is that the quality of the pH definition – the precision by which p*K* values used in the calculations are given and the precision of experimental p*I* values in these cases – will limit the possibilities to verify polypeptide compostion based on the experimental p*I* value.

Statistical comparison of experimental and calculated p*I* values was done using the *t*-test for dependent samples and normality of the discrepancies was estimated by probability plots. For the 36 proteins, the *p*-level is 0.0021, indicating that a result like this is unlikely to be a chance effect and must be assumed to represent a real difference. After correction for the most likely *N*-terminal configuration, the *p*-level is 0.043 and cannot be accepted as representing the same population since the *p*-level is less than 0.05 – the traditional *p*-limit of statistical significance. For the 18 proteins with a known or very likely *N*-terminal configuration the *t*-test gave a *p*-level of 0.49, which verifies that the experimental and calculated p*I* values are not significantly different.

Besides showing that p*I* values for denatured proteins with known compositions can be calculated with a high degree of precision from average p*K* values, the results also provide strong support for the notion that *N*-terminal blockage heavily depends on the nature of the *N*-terminal groups [26]. The results seem to indicate that with *N*-terminals other than M, S and A, only a few proteins have blocked *N*-terminals (1 out of 10 proteins in the present study), while it can be inferred from the data presented in Table 2 that a majority of the proteins with M, S and A as *N*-terminal are blocked. After correction for the effect of suspected *N*-terminal blockage there is only one protein (nucleolar protein B23) out of the 36 used in this study, which, in spite of a high buffer capacity, has a marked difference of 0.11 pH units between predicted and determined p*I* values (Fig. 4B); this corresponds to 3 charge units due to the high buffer capacity of this protein. This discrepancy in p*I* prediction and calculation of net charge at the p*I* is probably not due to deficiencies in the database information but instead reflects a shortcoming of the model used for p*I* calculations. Nucleolar protein B23 contains a domain extremely rich in aspartic and glutamic acid residues (Table 4), in which 26 out of 28 amino acid residues from position 161 to 188 are either a D or an E. A calculation based on the use of average p*K* values uninfluenced by the charged neighboring amino acid residues cannot be expected to correctly describe the p*I* value with almost half of the acidic groups packed together into a highly negatively charged region. This limitation caused by calculations based on average p*K* values does not severely limit the usefulness of the approach since a search through Swiss-Prot shows that this type of D/E-rich motif is uncommon, and the existence of a highly charged region is immediately apparent upon inspection of the amino acid sequence.

Table 4. Amino acid sequence of nucleolar phosphoprotein B23

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

The quality of the information available in databases, especially concerning posttranslational modifications, is a major problem when the data is to be used for p*I* predictions. The *p*-level of 0.043 found for all 36 proteins after correction for *N*-acetylation, shows that this problem is not only limited to *N*-terminal blockage and the very good agreement found for the eighteen polypeptides, with assumingly correctly described *N*-terminal (Fig. 4C), must be regarded as an exception from this point of view. *N*-Terminal blockage is generally the main problem in relation to p*I* predictions for eukaryotic proteins. Of the 36 keratinocyte proteins analyzed, 18–20 are suspected to be *N*-terminally blocked (6 proteins blocked according to Swiss-Prot, 12 proteins with M, S or A as *N*-terminal and assumingly blocked based on the calculated charge, and two proteins, involucrin and nucleolar protein B23, with M as *N*-terminal for which the data does not allow any conclusion). This is in reasonable agreement with the conclusions based on the *N*-terminal sequencing data derived in connection with 2-D gel electrophoresis. *N*-terminal blockage can be suspected for 17–19 of the 26 proteins with M, S or A as *N*-terminal, while only 1 in 10 proteins with other *N*-terminal groups are blocked. The information that the frequency of *N*-terminal blockage is strongly related to the nature of the *N*-terminal group will be of some help in connection with p*I* predictions based on database information. However, without information from other sources, an uncertainty will always remain as to whether the *N*-terminal charge should be included in the p*I* calculation.

4 Concluding remarks

The data presented here lays the foundation for comparing 2-D gel protein maps of different cell types generated with nonlinear, wide-range IPGs in the first dimension. The focusing positions of 41 polypeptides common to most human cell types have been described in a pH scale that allows focusing positions to be predicted with a high degree of accuracy, provided that the composition of the polypeptides are known and that information on posttranslational modifications are available. For polypeptides with a very high buffer capacity, the limiting factor is the precision with which experimental pH values can be determined rather than the precision of the calculations. Possible deficiencies in the pH scale description of the variation of the hydrogen ion activity has, at least at the present state, no consequences for its practical use. The major limitation in connection with predictions of focusing positions from polypeptide compositions is the quality of existing data on protein compositions, especially concerning posttranslational modifications. Amino acid sequences have been reasonably easy to obtain, while posttranslational modifications have been difficult and work-intensive to determine. Recent developments in the field of mass spectrometry are fast changing this situation and within the next years we can expect a surge in reliable data in this area. While awaiting this development, verification of correctness and completeness of available information on polypeptide composition can be provided by experimental p*I* values in a pH scale based on the p*I* values determined in this study. So far, our data cover the pH range below pH = 7.5. The basic pH range covered by NEPHGE as first dimension will be covered in forthcoming work.

Received December 29, 1993

5 References

[1] Gianazza, E., Astrua-Testori, S., Caccia, P., Giacon, P., Quaglia, L., Righetti, P. G., *Electrophoresis* 1986, *7*, 76–83.

[2] Görg, A., Postel, W., Günther, S., *Electrophoresis* 1988, *9*, 531–546.

[3] Hochstrasser, D. F., Frutiger, S., Paquet, N., Bairoch, A., Ravier, F., Pasquali, C., Sanchez, J.-C., Tissot, J.-D., Bjellqvist, B., Vargas, R., Appel, R. D., Hughes, G. J., *Electrophoresis* 1992, *13*, 992–1001

[4] *Immobiline DryStrip Kit for 2-D Electrophoresis: Instructions*, Pharmacia LKB Biotechnology AB, Uppsala 1993.

[5] Anderson, N. L., Hickman, B. J., *Anal. Biochem.* 1979, *93*, 312–320.

[6] Neidhardt, F. C., Appleby, D. A., Sankar, P., Hutton, M. E., Phillips, T. A., *Electrophoresis* 1989, *10*, 116–121.

[7] Rasmussen, H. H., Damme, J. V., Puype, M., Gesser, B., Celis, J. E., Vandekerckhove, J., *Electrophoresis* 1992, *13*, 960–969.

[8] Gianazza, E., Artoni, G., Righetti, P. G., *Electrophoresis* 1983, *4*, 321–326.

[9] Bjellqvist, B., Hughes, G. J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J.-C., Frutiger, S., Hochstrasser, D. F., *Electrophoresis* 1993, *14*, 1023–1031.

[10] Bjellqvist, B., Pasquali, C., Ravier, C., Sanchez, J.-C., Hochstrasser, D. F., *Electrophoresis* 1993, *14*, 1357–1365

[11] O'Farrell, P. H., *J. Biol. Chem.* 1975, *250*, 4007–4021

[12] Görg, A., *Biochem. Soc. Transactions* 1993, *21*, 130–132.

[13] Hanash, S. M., Strahler, J. R., Neel, J. V., Hailat, N., Melhem, R., Keim, D., Zhu, X. X., Wagner, D., Gage, D. A., Watson, J. T., *Proc. Natl. Acad. Sci. USA* 1991, *88*, 5709–5713.

[14] Görg, A., Postel, W., Friedrich, C., Kuick, R., Strahler, J. R., Hanash, S. M., *Electrophoresis* 1991, *12*, 653–655

[15] Celis, J. E., Rasmussen, H. H., Olsen, E., Madsen, P., Leffers, H., Honoré, B., Dejgaard, K., Gromov, P., Hoffmann, H. J., Nielsen, M., Vassilev, A., Vintermyr, O., Hao, J., Celis, A., Basse, B., Lauridsen, J. B., Ratz, G. P., Andersen, A. H., Walbum, E., Kjærgaard, I., Puype, M., Van Damme, J., Delay, B., Vandekerckhove, J., *Electrophoresis* 1993, *14*, 1091–1198.

[16] Celis, J. E., Madsen, P., Rasmussen, H. H., Leffers, H., Honoré, B., Gesser, B., Dejgaard, K., Olsen, E., Magnusson, N., Kiil, J., Celis, A., Lauridsen, J. B., Basse, B., Ratz, G. P., Andersen, A., Walbum, E., Brandstrup, B., Pedersen, P. S., Brandt, N. J., Puype, M., Van Damme, J., Vandekerckhove, J., *Electrophoresis* 1991, *12*, 802–872.

[17] Bjellqvist, B., Ek, K., Righetti, P. G., Gianazza, E., Görg, A., Postel, W., Westermeier, R., *J. Biochem. Biophys. Methods* 1982, *6*, 317–333.

[18] Bairoch, A., Boeckman, B., *Nucleic Acids Res.* 1991, *19*, 2247–2249

[19] Honoré, B., Madsen, P., Basse, B., Andersen, A., Walbum, E., Celis, J. E., Leffers, H., *Nucleic Acids Res.* 1990, *18*, 6692.

[20] Altland, K., *Electrophoresis* 1990, *11*, 140–147.

[21] Perrin, D. D., Dempsey, B., Serjeant, E. P., *pKa Predictions for Organic Acids and Bases*, Chapman and Hall Ltd., London 1981.

[22] Perrin, D. D., *Dissociation Constants of Organic Bases in Aqueous Solutions*, Butterworths, London 1965.

[23] Perrin, D. D., *Dissociation Constants of Organic Bases in Aqueous Solutions*, Supplement 1972, Butterworths, London 1972.

[24] Altland, K., Becher, P., Rossman, U., Bjellqvist, B., *Electrophoresis* 1988, 9, 474–485.

[25] Brown, J. L., Roberts, W. K., *J. Biol. Chem.* 1976, *251*, 1009–1014.

[26] Persson, B., Flinta, C., Heine, G., Jörnvall, H., *Eur. J. Biochem.* 1985, *152*, 523–527.